Australian Government
Department of Defence
Defence Science and
Technology Organisation

Speaker Localisation Using Time Difference of Arrival

*Derek Zong Thai, *MatthewTrinkle, Ahmad Hashemi-Sakhtsari and Tim Pattison*

Command, Control, Communication and Intelligence Division
Defence Science and Technology Organisation

***School of Electrical and Electronic Engineering**
The University of Adelaide

DSTO-TR-2126

ABSTRACT

This report describes the research and development of speaker localisation to locate the position of a person speaking. Two closed-form localisation techniques were analysed, the first was developed by Schau and Robinson (1987) based on spherical intersection and the other developed by Chan and Ho (1994). Both techniques are based on time difference of arrival measurements. Accordingly three time delay estimators, namely cross-correlation, generalised cross-correlation, and an eigenvalue decomposition based algorithm were analysed. The implementation of the algorithms in Matlab and the results from the analyses are discussed.

RELEASE LIMITATION

Approved for Public Release

Published by

Command and Control Division
Command, Control, Communication and Intelligence Division
Defence Science and Technology Organisation
PO Box 1500
Edinburgh South Australia 5111 Australia

Telephone: (08) 8259 5555
Fax: (08) 8259 6567

AR 014-178
April 2008

Conditions of Release and Disposal

Speaker Localisation using Time Difference of Arrival

Executive Summary

Speaker localisation is defined as the ability to automatically locate the position of a person speaking in a room. A localisation algorithm was implemented and tested for use in the Intense Collaboration Space (ICS) smart room environment at DSTO, Edinburgh.

Time Difference of Arrival (TDOA) was the localisation method chosen. TDOA localisation is based on two parts: a time-delay estimator and a localisation estimator. Two closed-form localisation algorithms, the Spherical Intersection method (Schau and Robinson 1987) and a closed-form method by Chan and Ho (1994) were chosen for implementation due to their simplicity and speed. Both algorithms were found to be robust against small time delay errors, with the Chan and Ho estimator giving a smaller variance.

Three time-delay estimators, the ordinary cross-correlation, the generalised cross-correlation using phase transform (GCC-PHAT) and an eigenvalue decomposition (ED) algorithm were studied. GCC-PHAT and ED algorithms were implemented in MATLAB and were compared with MATLAB's cross-correlation function.

Experiments were conducted in the ICS to obtain data for testing purposes. Noise signals of varying sound levels were played out through a computer speaker and recorded using the wall microphones that exist in the room. The recorded data also included the background noise and reflected sounds from the room.

The results of the tests showed that GCC-PHAT and ED methods worked much better than the ordinary cross-correlation in the presence of existing background noise. GCC-PHAT is faster than ED in terms of computation requirement and would therefore be more suited for real-time use.

Authors

Derek Zong Thai

Command, Control, Communication and Intelligence Division , DSTO Edinburgh

Derek Thai, graduated in 2005 from B.Engineering (Computer Systems)/B.Economics at The University of Adelaide with a final year Engineering project of adaptive noise cancellation sponsored by DSTO. In 2006 Derek accepted a GILES scholarship at DSTO to work on a project of speaker localisation and to study a Masters degree in Project Management.

Matthew Trinkle

School of Electrical and Electronic Engineering
The University of Adelaide

Matthew Trinkle is a Research Fellow in the School of Electrical and Electronic Engineering of the University of Adelaide. His area of expertise and interest is in Signal Processing including audio processing, speaker localisation and tracking and automatic noise and echo cancellation.

Ahmad Hashemi-Sakhtsari

Command, Control, Communication and Intelligence Division , DSTO Edinburgh

Ahmad Hashemi-Sakhtsari is a Defence Scientist in C3I Division in DSTO Edinburgh. His area of research interest is in speech and language technology, including human computer interaction and speech processing.

Tim Pattison

Command, Control, Communication and Intelligence Division , DSTO Edinburgh

Tim Pattison obtained a BSc in Mathematical Science, and a BE (Hons) and PhD in Electrical and Electronic Engineering, all from the University of Adelaide, as well as a Graduate Certificate in Management from the University of South Australia.

Since joining DSTO in 1987, Tim has worked in the Communications, Information Technology, Command and Control, and C3I Divisions of DSTO, most recently as Head of the newly-formed Human Interaction Capabilities Discipline. In 2004, he was awarded a DCJOPS Commendation for his work in support of Theatre intelligence capabilities.

His recent research interests include: information visualisation, text and data mining, and speech and language technologies.

Contents

1. Introduction

The Command, Control, Communications and Intelligence (C3I) Division in DSTO is conducting research into networked smart-room environments – known as Livespaces (Weber, 2007) – for distributed collaborative planning. Knowledge of the location of its occupants would allow a Livespace to better support them by: steering microphone arrays to improve the quality of audio pickup for recording, communication and transcription; enhancing the separation – and hence recognition – of speech from concurrent speakers; steering and zooming cameras for enhanced face and gesture recognition during video teleconferencing; adapting the local visual and auditory presentation of information; and providing explicit or implicit awareness of occupant location to participants at remote sites.. Whilst acknowledging that the identification of occupants would afford additional benefits, this report focuses solely on their localisation.

Localisation methods can be broadly classified according to whether or not the room occupants contribute to their localisation by carrying or wearing equipment which assists that localisation. Such equipment might for example include: a GPS-like device which calculates position using signals from beacons of known location; an emitter or transponder whose reception by spatially disparate sensors within the room is used to calculate position; or even tabs of reflective material which enhance contrast and hence detection by video processors. Encumbrance of personnel should be minimised for a variety of reasons, including: impaired mobility due to weight, size or tethering of equipment; time and effort required to don, activate and later shed the equipment; breakage and loss of shared equipment; the potential for an individual to become overburdened with a proliferation of encumbrances such as stereo glasses, a personal microphone, identification tag and personal digital assistant; and the need to power portable devices which are active.

A variety of sensing modalities, including video, audio, infrared and ultrasound, could contribute to the detection, localisation and tracking of an unencumbered occupant of a room. Fusion of multiple modalities would further provide for more accurate and reliable localisation (see e.g. Gatica-Perez et al., 2006, on fusing video and audio), albeit at the expense of additional sensors and processing. In order to decide on a suitable mix of modalities for a particular application/requirement, we need to understand the physical and technical limitations and capabilities of each: occlusion, spatial resolution, background noise and clutter, temporal discontinuity of the stimulus, scalability to multiple targets, sensor and processing cost.

This report measures the accuracy of techniques available in localising and tracking individual participants through their speech signals received at multiple microphones whose position within the room are known, or can be calculated through prior microphone calibration techniques. In answering the question as to why use speech, it is advocated that meeting environments are likely to already have a number of microphones and hence may not need additional sensors and a great deal of processing. In such environments signal processing equipment e.g. beamforming microphone arrays may already be available for co-channel speaker separation, noise or echo cancellation that

would improve the performance of speaker localisation and tracking algorithms. As with audio beamforming, the complexity of measurements reported in this paper is confined to the current speaker, since this is required in applications such video zoom and pan for VTC. Nevertheless intermittent nature of speech remains an inherent limitation in speech processing.

The bulk of the work described in this report was undertaken by Derek Thai during his placement at DSTO Edinburgh under the Graduate Industry-Linked Entrepreneur Scheme (GILES). Derek was co-supervised by Ahmad Hashemi-Sakhtsari (DSTO) and Matthew Trinkle (University of Adelaide), both of whom made substantial contributions to the report following the completion of Derek's placement. In acknowledgement of his significant editorial input, Tim Pattison was subsequently invited to become an author.

1.1 Aim

The aim of this project is to describe the design of a real-time speaker localisation system for use in smart workspace environments for collaborative activities. This system is intended to be implemented in the Intense Collaboration Space (ICS) Laboratory smart room at DSTO.

The main factors determining the accuracy of the final system are expected to be background noise in the room and reverberation. The ICS Laboratory has felt-covered walls which help reduce reverberations to a level that allows localisation results to be obtained relatively consistently without requiring special techniques for reducing the effects of reverberation. However, the computer rack in the ICS room appears to be producing a significant amount of background noise at low frequencies that includes the speech band. This noise source is a potential problem as it is also localised and will not necessarily be much reduced by cross-correlation processing because it is partially correlated similarly to the desired speech signal.

2. Background

There are many methods used for speaker localisation. These can be loosely divided into three categories: direct methods, high-resolution spectral estimation-based methods, and time difference of arrival methods (Brandstein and Silverman 1997).

2.1 Direct Methods

Direct methods derive the estimate directly from the filtered, weighted and summed signal data received at the sensors. These are steered beamformer-based methods, which steer an array of microphones to search for peaks in source output power (Brandstein and Silverman 1997). These direct methods can use *a priori* knowledge of spectral information of the signal and noise to improve their performance. The localisation functions are computationally complex, and so are impractical for real-time localisation. Direct methods include maximum likelihood (ML) estimators (Bangs and Shultheis 1973; Hahn and

Tretter, 1973; Hahn 1975; Carter 1977), standard iterative optimisation methods (Wax and Kailath 1983) and correlation-based estimators (Silverman and Kirtman 1992).

2.2 High-resolution Spectral Estimation-based Methods

High-resolution spectral estimation-based methods involve estimating the spatiospectral correlation matrix from the data received at the sensors (Brandstein and Silverman 1997). These methods are less robust to source and sensor modelling errors compared with beamforming methods and assume conditions that do not exist in the real world, such as an exact knowledge of sensors, ideal source radiators and uniform sensor channel characteristics (Brandstein and Silverman 1997). Although the computational complexity is not as intense as beamformer-based methods, they are still considerable and typically increase as additional methods are used to improve performance. Examples of these types of methods include autoregressive (AR) modelling, minimum variance (MV) spectral estimation and eigenvalue based techniques (Haykin 1991; Johnson and Dudgeon 1993).

2.3 Time Difference of Arrival Methods

Time difference of arrival (TDOA) estimators are based on a two-step method. Initially, relative time delays for each source are estimated from the received data. These time delays are then used to generate hyperbolic curves which intersect at the source location estimate as shown in Figure 1.

TDOA methods are usually based on the generalised-cross-correlation (GCC) to obtain time delays (Knapp and Carter, 1976), especially using the Phase Transform (GCC-PHAT) method (Macho et al. 2005). These TDOA methods are the most popular of the three due to their relatively low computational complexity.

The two stage process, although reducing the complexity, is sub-optimal as a location estimator. It is also difficult for this method to work with multi-source scenarios since algorithms assume a single-source model (Brandstein and Silverman, 1997).

Two main types of TDOA methods are ML estimators and closed-form estimators. ML methods, such as the Taylor Series method, are iterative and computational complexity can be high, although not as high as direct methods or spectral estimation-based methods. Closed-form methods, such as the Chan and Ho method (Chan and Ho 1994), attempt to linearise the non-linear estimation process. They are computationally fast and usually suffer from little detriment in performance. Due to their sub-optimal nature, closed-form estimators can be used as an intermediate step by providing a starting point for more computationally complex methods (Brandstein and Silverman 1997).

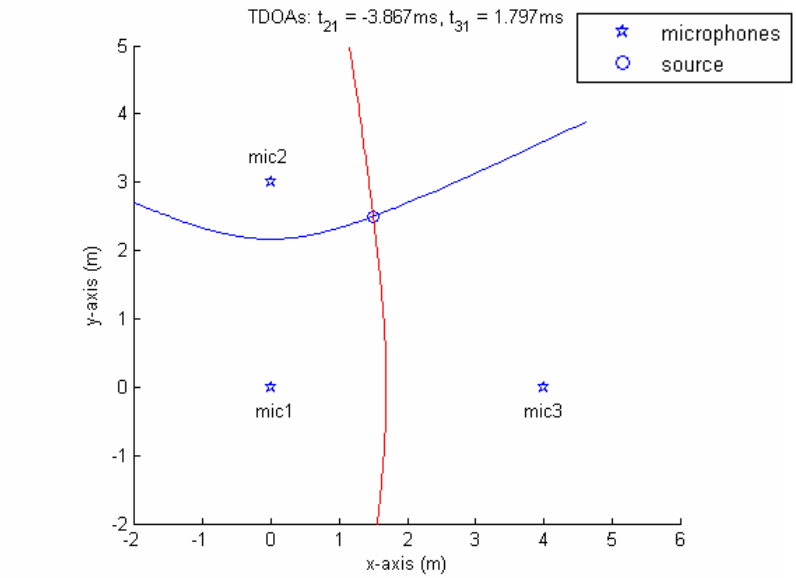

Figure 1 Intersection of two hyperbolas created from TDOA estimates provides the position of the speaker. A minimum of three microphones is required for 2D localisation. In this figure microphones 2 and 3 are referenced to microphone 1.

3. Project Overview

The work reported here is part of an ambitious larger modular software system that aims to integrate real-time audio and video recording and transcription modules to achieve data fusion for workflow monitoring. Ultimately an integrated environment for speaker localisation and tracking will be set up to collect meeting participants' data using audio, video and text on a DVD. As well as meetings and interviews, the system could be tailored to be used for training, teaching, analysis, scriptwriting and developing protocols.

This project focuses on using TDOA for speaker localisation in a smart room environment. TDOA localisation has the most efficient computational complexity and so is more suited to real-time applications than other localisation methods. The localisation system is intended to be implemented in the ICS Laboratory at DSTO. The ICS Laboratory is a smart collaboration space that is part of the LiveSpaces initiative. It employs a computer controlled adaptive environment with communication and interaction facilities such as notebook/tablet computers, screen projectors, cameras, microphones and multiple writing surfaces for hand diagrams and writings which can be digitally captured (see Figure 2).

Figure 2. ICS Laboratory at DSTO.

There are three main parts to this project. The first part is the research into TDOA methods and their requirements (Section 4). The second part is the implementation of the selected algorithms in MATLAB for testing and evaluation purposes (Section 5). The third part is the conduct of experiments using the implemented algorithms to determine which is more effective for real-time speaker localisation (Section 6).

3.1 Background Research

Research was conducted on two fronts: background research into TDOA localisation algorithms, and research into time-delay estimators that are used for TDOA localisation.

Two closed-form TDOA methods were chosen for this project: Spherical Intersection (SX) method (Schau and Robinson 1987) and another method proposed by Chan and Ho (1994) (dubbed the Ho-Chan method). The main reason for this choice was that these methods are very fast and efficient when compared to other localisation methods, and hence most useful in a real-time system. A real time system would be required if a camera were to be used to track a speaker. Other reasons include the straightforwardness of the implementation of this method and the fact that closed-form estimators do not require an initial estimate. Previous work done by Rice (2004) and Pang[1] was based on the SX method.

Three time-delay estimators were researched for this project: the ordinary cross-correlation, the generalised cross-correlation and the eigenvalue decomposition algorithm.

[1] Daniel Pang researched the SX method while conducting a project for DSTO based on speaker localisation in 2005.

3.2 Implementation of Algorithms

The localisation system was implemented in three parts: the localisation algorithm, the time-delay estimators, and the iterator.

The localisation algorithm reads in the relative time-delays between microphones and the microphone positions as inputs. Based on this information it produces the source position estimate. The time-delay estimators produce the relative time-delays for the localisation algorithm. All three algorithms mentioned in section 3.1 were implemented for comparison.

The iterator is a program that includes a time-delay estimator and the localisation algorithm. It repeats the localisation process multiple times to track the sound source. There are two versions of the iterator, the first reads in pre-recorded computer audio files as input, the second reads from the computer audio I/O card so that data received at the microphones can be used directly. This very simple tracking method was implemented due to time constraints. In future more sophisticated tracking algorithms such as the Kalman Filter or Particle Filter should be used.

3.3 Experiments and Analysis

A set of experiments were conducted in the ICS Laboratory. Computer generated noise signals of different bandwidths (see Table 1) were emitted from a computer speaker in the room and were recorded using microphones attached to the walls. The recorded audio data included background noise and reverberation that exist in the room, and were used to test the effectiveness of the time-delay estimators.

4. Research

Localisation techniques based on TDOA estimates are a two part process. First the relative time-delays between the receivers are estimated from the received data, and then the localisation algorithm uses the time-delays to estimate the position of the source. The research of the algorithms hence is divided into two main sections: the localisation algorithms and the time-delay estimation methods.

4.1 Localisation Algorithms

The SX method by Schau and Robinson (1987) and the Ho-Chan method by Chan and Ho (1994) were the two TDOA localisation methods investigated in this project.

The basis behind the two methods is similar in that both methods rely on minimising a least squares (LS) estimator. The SX method utilises the relationship in (1) between the source coordinates (x, y) and distance r_i when finding the coordinate estimates:

$$(x_i - x)^2 + (y_i - y)^2 = r_i^2 \qquad (1)$$

where (x_i, y_i) are the coordinates of sensor i.

The Ho-Chan method initially assumes that the three variables x, y and r_i are independent and uses LS to find an initial estimate. It then puts the relationship in (1) into the calculation combined with the initial estimate to produce the final estimate. This results in estimates that are optimum in a 'Least Squares' sense whereas the SX method does not (Chan and Ho 1994).

4.1.1 Spherical Intersection (SX) Method

The SX method uses an array of at least three sensors of which one is selected to be the reference sensor. Without any loss of generalisation, let sensor 1 be the reference sensor. Thus for $i = 1$ the equation (1) becomes

$$(x_1 - x)^2 + (y_1 - y)^2 = r_1^2 . \qquad (2)$$

Let $X = \begin{bmatrix} (x_2 - x_1) & (y_2 - y_1) \\ (x_3 - x_1) & (y_3 - y_1) \\ \vdots & \vdots \\ (x_N - x_1) & (y_N - y_1) \end{bmatrix}$, $\quad K_i = x_i^2 + y_i^2$ for $i = 1$ to N

where N is the number of sensors.

Subtracting (2) from (1) where $i = 2$ to N gives a set of equations which can be written in the form

$$\begin{bmatrix} x \\ y \end{bmatrix} = C r_1 + D \qquad (3)$$

where

$$C = \left(X^T X\right)^{-1} X^T \begin{bmatrix} -r_{2,1} \\ -r_{3,1} \\ \vdots \\ -r_{N,1} \end{bmatrix}, \qquad D = \left(X^T X\right)^{-1} X^T \left(\frac{1}{2}\right) \begin{bmatrix} -r_{2,1}^2 + K_2 - K_1 \\ -r_{3,1}^2 + K_3 - K_1 \\ \vdots \\ -r_{N,1}^2 + K_N - K_1 \end{bmatrix}.$$

Where $r_{i,j}$ is the TDOA distance between microphones j and i.

Substituting (2) into (3) gives a quadratic equation of the form

$$\alpha r_1^2 + \beta r_1 + \chi = 0 . \qquad (4)$$

For a complete derivation see Appendix A.1.

Solving the quadratic in (4) produces two results for r_1. The correct result can then be put back into (3) to solve for (x, y). However choosing the correct result can be problem. There are four possible solutions that can occur in the quadratic: two negative solutions, two positive solutions, one positive and one negative solution or two complex solutions.

Pang researched this problem and found that a correct result exists when the quadratic roots are either both positive or one is positive and one is negative. When the roots are both positive the smaller root should be chosen. When there is one positive and one negative root the positive root should be chosen as r_1 is distance measure and cannot be negative. When the other two types of roots occur (both negative or complex) it means that there is no solution to the problem and the source cannot be estimated.

Simulations were conducted based on this selection process to determine its accuracy. A virtual rectangular room of 4 x 4 m was used in the simulations (although any sized rectangular room would suffice). Four microphone sensors were placed in the room in two configurations: the first configuration had a microphone in each corner of the room; the second configuration had a microphone in the middle of each wall.

Figures 3-7 show the results of configuration 1. The microphones are indicated by the red and black dots; the red dot is the reference microphone. Figure 3 shows the coordinates found when there is one positive and one negative root. Figure 4 shows the coordinates found when there are two positive roots. Figure 5 shows the combined coordinates.

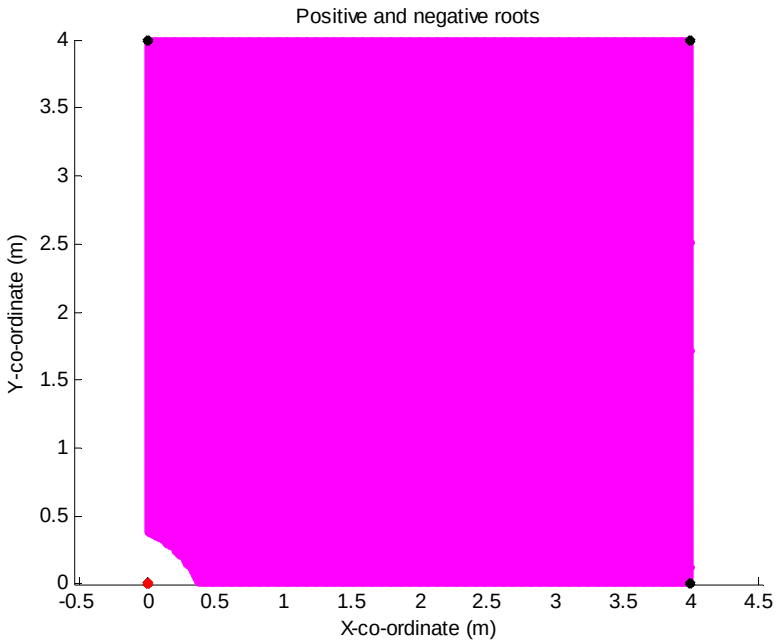

Figure 3. Microphone configuration 1; the shaded area indicates where the quadratic has one positive and one negative root.

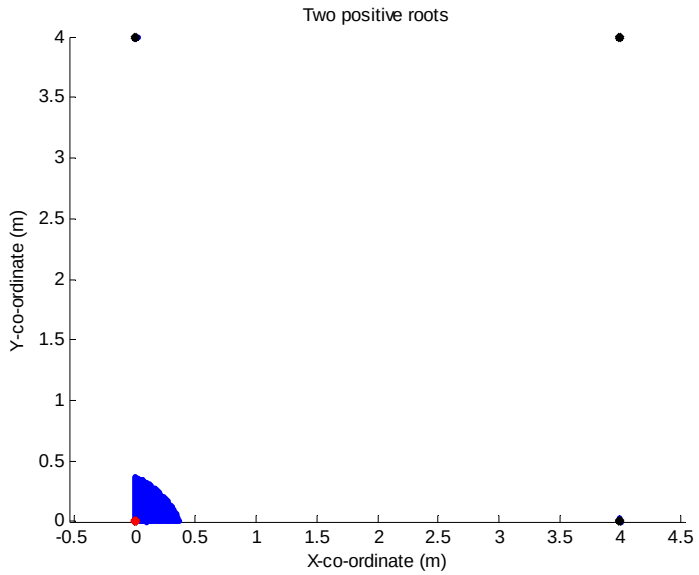

Figure 4. Microphone configuration 1; the shaded area indicates where the quadratic has two positive roots.

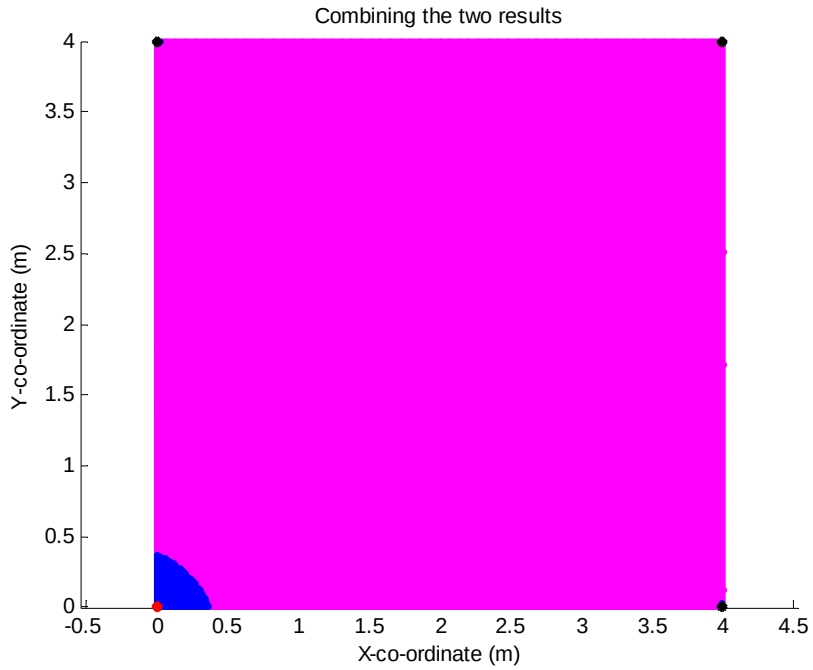

Figure 5. Microphone configuration 1; combining the areas shown in Figures 1 and 2. The pink/magenta area is where one positive root and one negative root occurred; the blue area is where two positive roots occurred.

As can be seen in Figure 5 the whole room can be covered by this selection process. There is a distinct line between the areas where two positive roots occur and where one positive

root and one negative root occur. To see whether there are irregularities along that line, standard deviation plots were created using the same configuration.

Note how there is no distinct break line that occurs in Figure 5. This indicates that the selection process is correct and the solution found is unique. Figures 6 and 7 show the standard deviation plots of the estimates when random white Gaussian noise with a fixed variance was added to the TDOA estimates.

The transition between two positive real roots, and one positive plus negative root, was studied in more detail as this situation gives rise to a potential discontinuity in r_1. However it was found that the smaller of the two positive roots remained continuous at the transition, while the larger positive root was discontinues and became negative at the transition. As the smaller of positive root is chosen as the solution for r_1, it was continuous.

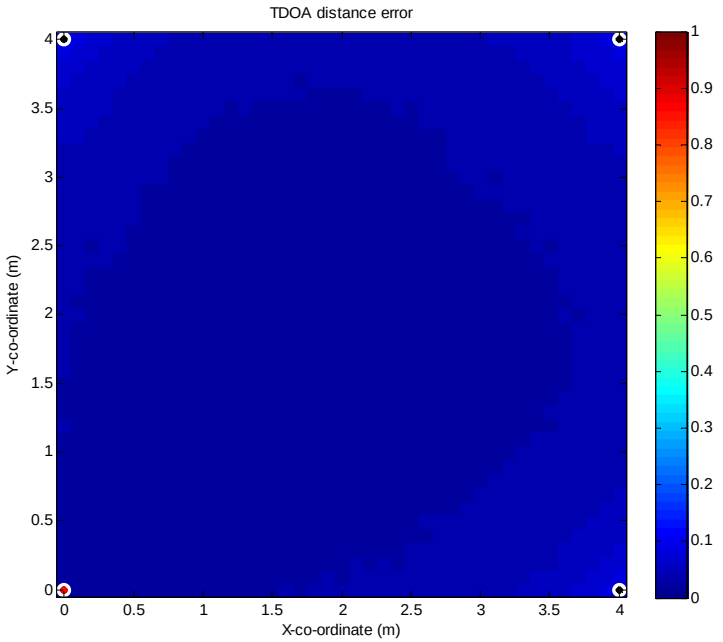

Figure 6. Microphone configuration 1; plot of standard deviation of coordinates over 500 samples when TDOA values had additive noise of variance 0.001 m². The colour bar on the right side shows the relative colour of the standard deviation between 0 and 1 m.

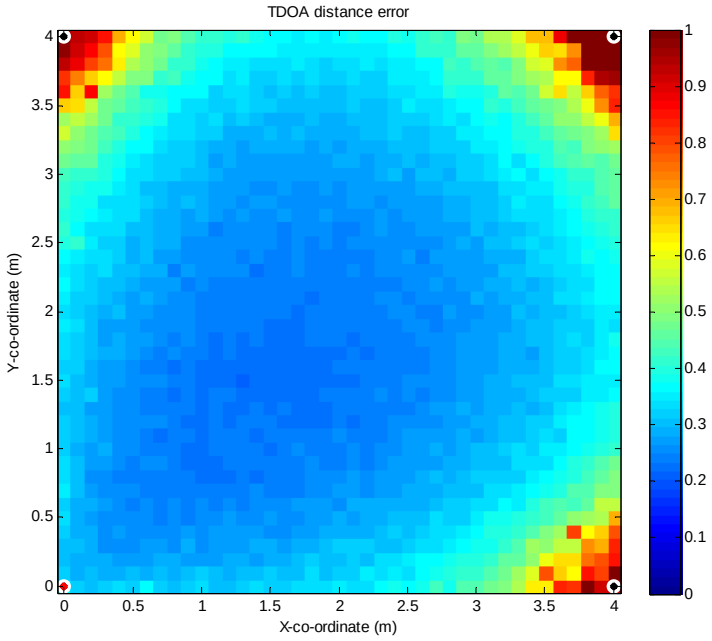

Figure 7. Microphone configuration 1; plot of standard deviation of coordinates over 500 samples when the TDOA values had additive noise of variance 0.1 m². The colour bar on the right side shows the relative colour of the standard deviation between 0 and 1 m.

Figures 8-12 show the same as Figures 3-7 but for the second configuration of microphones.

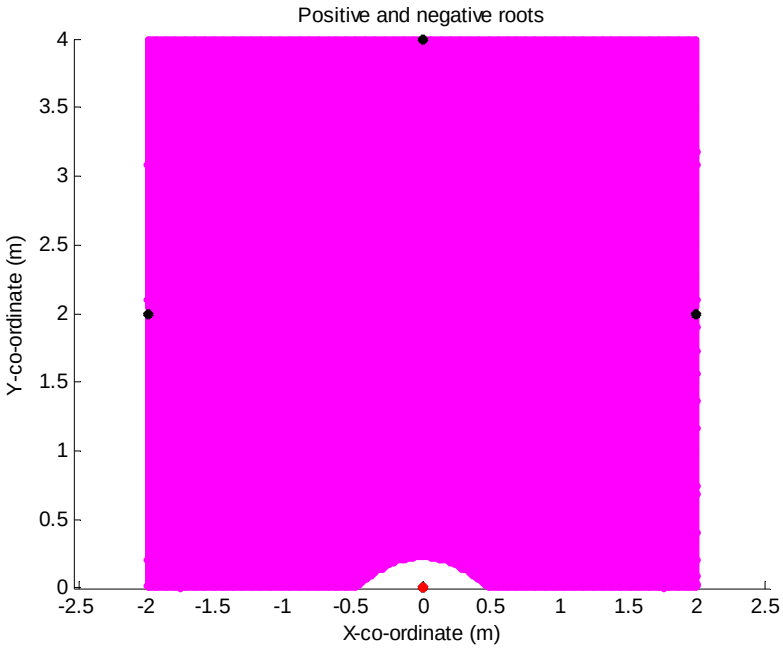

Figure 8. Microphone configuration 2; the shaded area indicates where the quadratic has one positive root and one negative root.

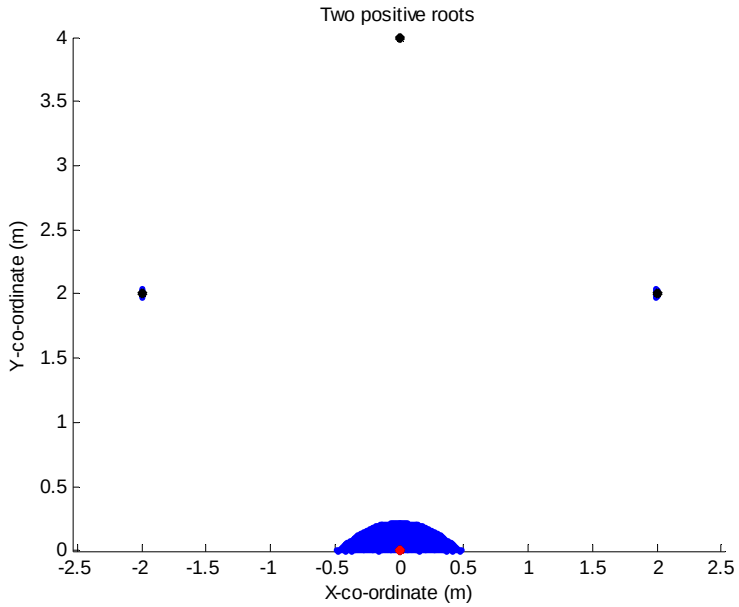

Figure 9. Microphone configuration 2; the shaded area indicates where the quadratic has two positive roots.

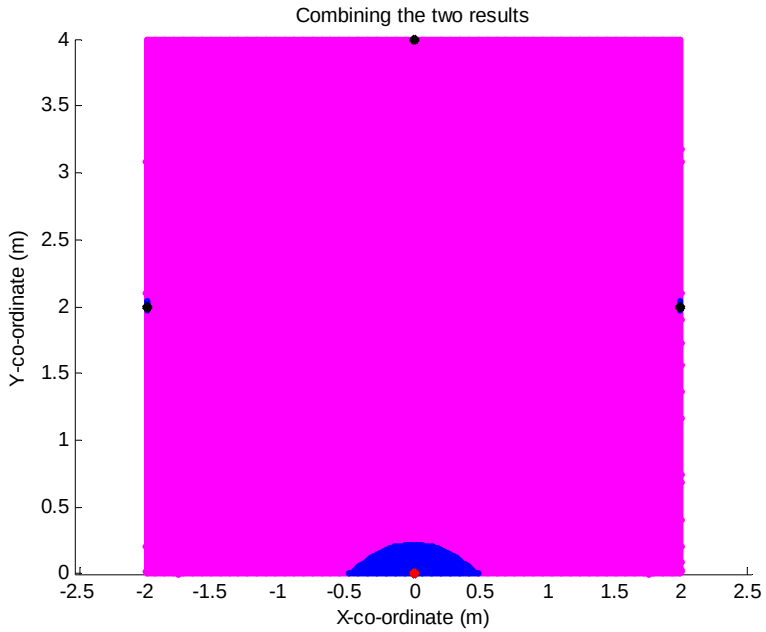

Figure 10. Microphone configuration 2; combining the areas shown in Figures 8 and 9. The pink/magenta area is where one positive root and one negative root occurred; the blue area is where two positive roots occurred.

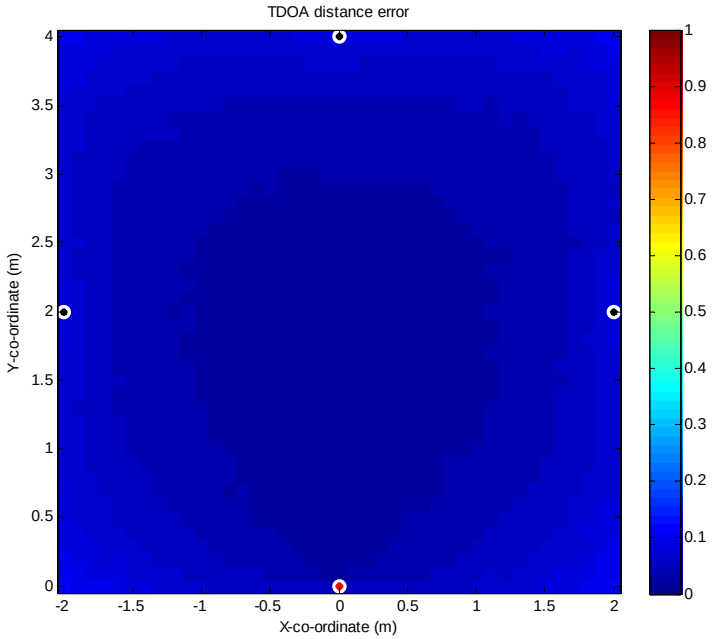

Figure 11. Microphone configuration 2; plot of standard deviation of coordinates over 500 samples when the TDOA values had additive noise of variance 0.001 m². The colour bar on the right side shows the relative colour of the standard deviation between 0 and 1 m.

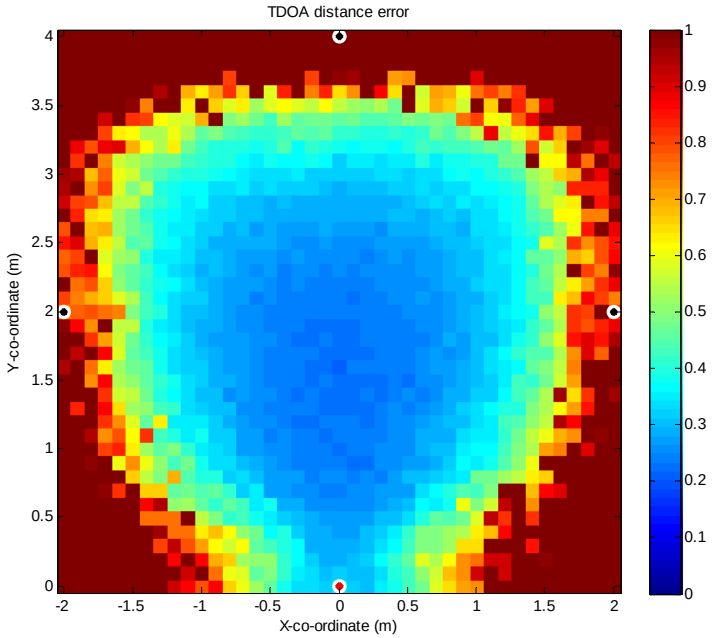

Figure 12. Microphone configuration 2; plot of standard deviation of coordinates over 500 samples when the TDOA values had additive noise of variance 0.1 m². The colour bar on the right side shows the relative colour of the standard deviation between 0 and 1 m.

Note that the array configuration in Figure 12 results in greater position variance than the configuration in Figure 7 especially in the corners of the room. This is expected to be due to the microphones being positioned further apart in Figure 7, thus enclosing a larger area and giving a better Geometric Dilution of Precision (GDOP) over a larger area.

4.1.2 Ho-Chan Method

This section gives the main steps and motivation of the Ho-Chan method. For a more detailed derivation as well as definition of all terms used in this discussion see Appendix A.

The Ho-Chan method starts with the same equations as the SX method and arrives at the same result as (3). By assuming that the variables x, y and r_1 are independent, the error vector ψ is defined as

$$\mathbf{h} - \mathbf{G}_a\mathbf{z}_a^0 = \psi \tag{5}$$

where

$$\mathbf{h} = -\frac{1}{2}B = \frac{1}{2}\begin{bmatrix} r_{2,1}^2 - K_2 + K_1 \\ r_{3,1}^2 - K_3 + K_1 \\ \vdots \\ r_{N,1}^2 - K_N + K_1 \end{bmatrix}, \qquad \mathbf{G}_a = -\begin{bmatrix} X & A \end{bmatrix} = -\begin{bmatrix} x_{2,1} & y_{2,1} & r_{2,1} \\ x_{3,1} & y_{3,1} & r_{3,1} \\ \vdots & \vdots & \vdots \\ x_{N,1} & y_{N,1} & r_{N,1} \end{bmatrix},$$

$$\mathbf{z}_a^0 = \begin{bmatrix} x^0 \\ y^0 \\ r_1^0 \end{bmatrix}, \qquad \psi = \begin{bmatrix} \psi_2 \\ \psi_3 \\ \vdots \\ \psi_N \end{bmatrix}.$$

$\mathbf{z}_a{}^0$ are the true values of $\mathbf{z}_a = \begin{bmatrix} x \\ y \\ r_1 \end{bmatrix}$ which correspond to the estimated source location (x,y) and the estimated distance between the source and the reference microphone r_1.

When the true TDOA distances are available $\psi = 0$. In practice this is not the case. The Ho-Chan method involves a Least Square (LS) calculation to minimise ψ:

$$\mathbf{z}_a = (\mathbf{G}_a{}^T\mathbf{\Psi}^{-1}\mathbf{G}_a)^{-1}\mathbf{G}_a{}^T\mathbf{\Psi}^{-1}\mathbf{h}. \tag{6}$$

$\mathbf{\Psi}$ is the covariance matrix of ψ and is given by

$$\mathbf{\Psi} = \mathbf{BQB} \tag{7}$$

where

$\mathbf{B} = \begin{bmatrix} r_2^0 & 0 & \cdots & 0 \\ 0 & r_3^0 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & r_N^0 \end{bmatrix}$, $\mathbf{Q}$ is the covariance matrix of the TDOA distances.

In practice $\mathbf{\Psi}$ is not known since $\mathbf{B}$ contains the true TDOA distances. The Ho-Chan method approximates (6) as

$$\mathbf{z}_a \approx (\mathbf{G}_a^T\mathbf{Q}^{-1}\mathbf{G}_a)^{-1}\mathbf{G}_a^T\mathbf{Q}^{-1}\mathbf{h}. \quad (8)$$

$\mathbf{B}$ is estimated from the initial estimates of $\mathbf{z}_a$ from (8) and is put into (6) to achieve a better estimate of $\mathbf{z}_a$. (6) can be iterated multiple times to produce an even better estimate, but simulations done by Chan and Ho (1994) show that one iteration is sufficient to produce accurate results.

The covariance of $\mathbf{z}_a$ is given by

$$\mathrm{cov}(\mathbf{z}_a) = (\mathbf{G}_a^{0T}\mathbf{\Psi}^{-1}\mathbf{G}_a^0)^{-1} \quad (9)$$

where $\mathbf{G}_a^0$ is identical to $\mathbf{G}_a$ except that it uses the exact values of $r_{2,1}$ .. $r_{N,1}$ rather than the estimated ones.
Let the elements of $\mathbf{z}_a$ be expressed as

$$\mathbf{z}_a = \begin{bmatrix} z_{a,1} \\ z_{a,2} \\ z_{a,3} \end{bmatrix} = \begin{bmatrix} x^0 + e_1 \\ y^0 + e_2 \\ r_1^0 + e_3 \end{bmatrix} \quad (10)$$

where e_1, e_2 and e_3 are estimation errors of $\mathbf{z}_a$.

Chan and Ho define another set of equations:

$$\mathbf{h}' - \mathbf{G}_a'\mathbf{z}_a'^0 = \psi' \quad (11)$$

where

$$\mathbf{h}' = \begin{bmatrix} (z_{a,1} + x_1)^2 \\ (z_{a,2} + y_1)^2 \\ z_{a,3}{}^2 \end{bmatrix}, \mathbf{G}_a' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{z}_a' = \begin{bmatrix} (x - x_1)^2 \\ (y - y_1)^2 \end{bmatrix}, \quad \psi' = \begin{bmatrix} \psi_1' \\ \psi_2' \\ \psi_3' \end{bmatrix}.$$

ψ' is the vector of inaccuracies in $\mathbf{z}_a$. The covariance matrix $\mathbf{\Psi}'$ of ψ' is

$$\mathbf{\Psi}' = \mathbf{E}[\psi'\psi'^{T}] = 4\mathbf{B}'\mathrm{cov}(\mathbf{z}_a)\mathbf{B}' \tag{12}$$

where

$$\mathbf{B}' = \begin{bmatrix} (x^0 - x_1) & 0 & 0 \\ 0 & (y^0 - y_1) & 0 \\ 0 & 0 & r_1^0 \end{bmatrix}.$$

$\mathbf{B}'$ can be approximated by using the values in $\mathbf{z}_a$. A second LS calculation is used to find $\mathbf{z}_a'$

$$\mathbf{z}_a' = (\mathbf{G}_a'^{T}\mathbf{\Psi}'^{-1}\mathbf{G}_a')^{-1}\mathbf{G}_a'^{T}\mathbf{\Psi}'^{-1}\mathbf{h}' \tag{13}$$

$\mathbf{z}_a'$ is an estimate of $(x - x_1)^2$ and $(y - y_1)^2$ and so a simple conversion will produce x and y estimates:

$$\mathbf{z}_p = \pm\sqrt{\mathbf{z}_a'} + \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}. \tag{14}$$

Since there are two possible solutions for a square root, the correct solution is the one that lies in the region of interest. This can be easily determined by looking at the sign of the initial $\mathbf{z}_a$ estimates. If one of the coordinates is close to zero the square root may become imaginary. In this case the imaginary component should be set to zero (Chan and Ho 1994).

To summarise, (8) is used to find an initial estimate of $\mathbf{z}_a$ which is used to estimate $\mathbf{B}$. (6) is then used to find more accurate estimates of $\mathbf{z}_a$ and can be repeated for multiple iterations. (13) is then used to put the relationship in (1) between x, y and r_1 into the calculation. The final coordinate estimate $\mathbf{z}_p$ is found using (14).

4.1.3 Comparison of Localisation Algorithms

Comparisons were made between the SX and Ho-Chan methods and the Cramér-Rao Lower Bound (CRLB) that gives the achievable minimum error. The CRLB sets a lower bound on the variance of any unbiased estimator. For derivation of the CRLB for TDOA localisation see Appendix A.3.

Simulations were done using two virtual rectangular rooms of size 4x4 m and 8x4 m. Six microphone sensors were placed on two opposite walls of the room as shown in Figure 13; the microphones are indicated by the red and black dots. White Gaussian noise of variances 0.01 m^2 and 0.0001 m^2 were added to the TDOA values calculated from each point in the room to generate standard deviation plots for the CRLB, SX method and Ho-Chan method. These plots are shown in Figures 21 to 64 in Appendix B.

Figures 13-15 show the standard deviation plots for one case where the room size is 8x4 m and the noise variance is 0.0001 m^2. The red dot is the reference microphone.

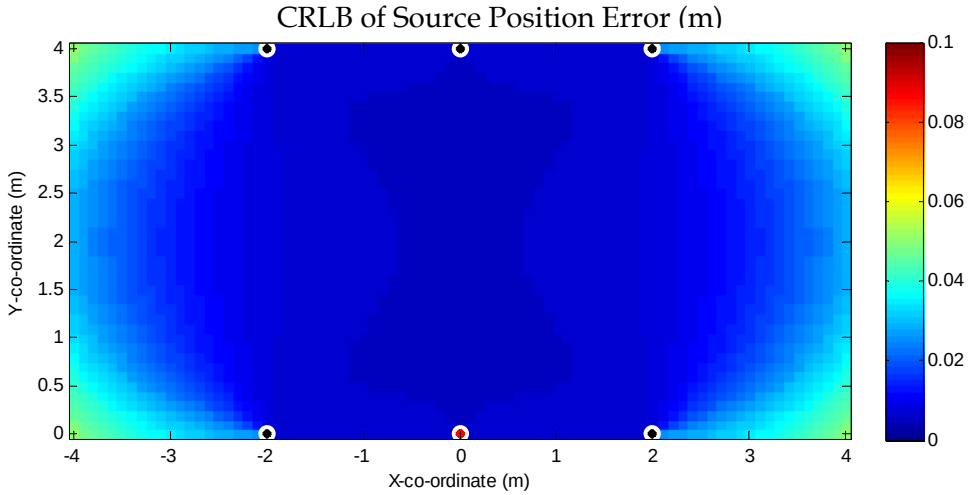

Figure 13. Standard deviation of CRLB position error when TDOA distances had additive noise of variance 0.0001 m^2; reference sensor in the middle.

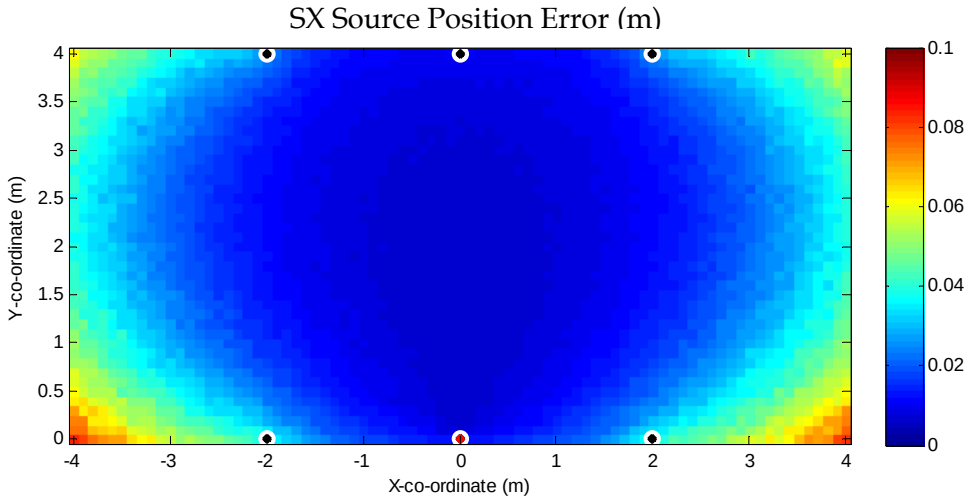

Figure 14. Standard deviation of position error using the SX method over 500 samples when TDOA distances had additive noise of variance 0.0001 m^2; reference sensor in the middle.

Ho Chan Source Position Error (m)

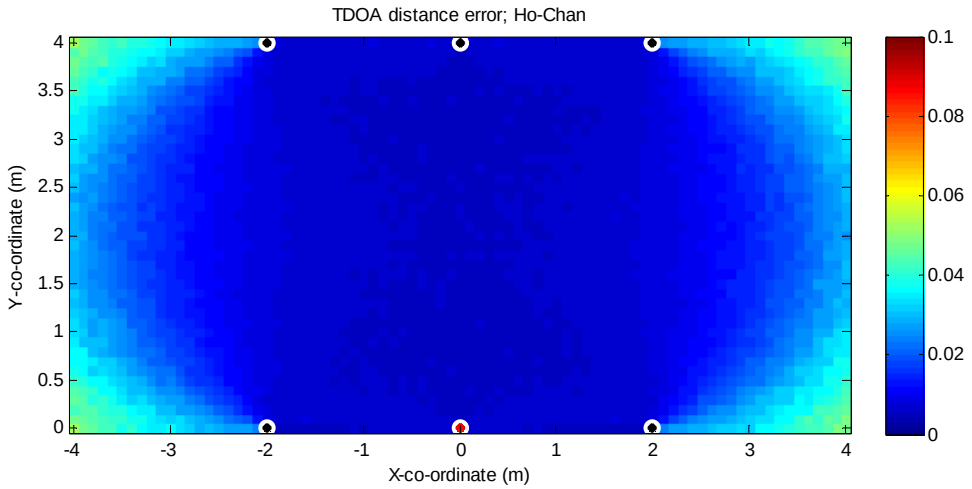

Figure 15. Standard deviation of position error using Ho-Chan method over 500 samples when TDOA distances had additive noise of variance 0.0001 m²; reference sensor in the middle.

By observation the Ho-Chan results resemble the CRLB much more closely than the SX results. To see this more accurately the CRLB was subtracted from the standard deviations of SX and Ho-Chan methods and are shown in Figures 16 and 17.

Note that Figures 16 and 17 are on a scale 10 times smaller than Figures 13-15. It is clear that the Ho-Chan method produce very accurate estimates provided that the input errors are small (in this case the error variance is 0.0001 m²).

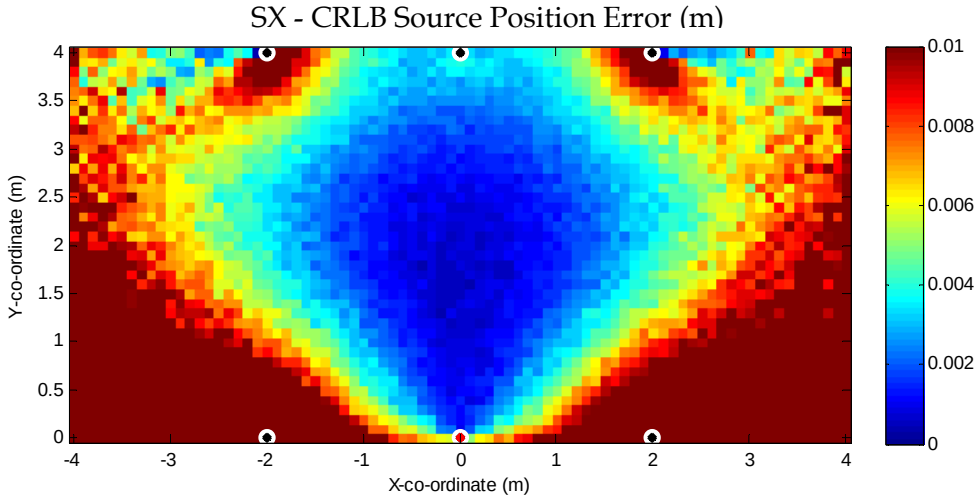

Figure 16. Plot of CRLB subtracted from SX standard deviations.

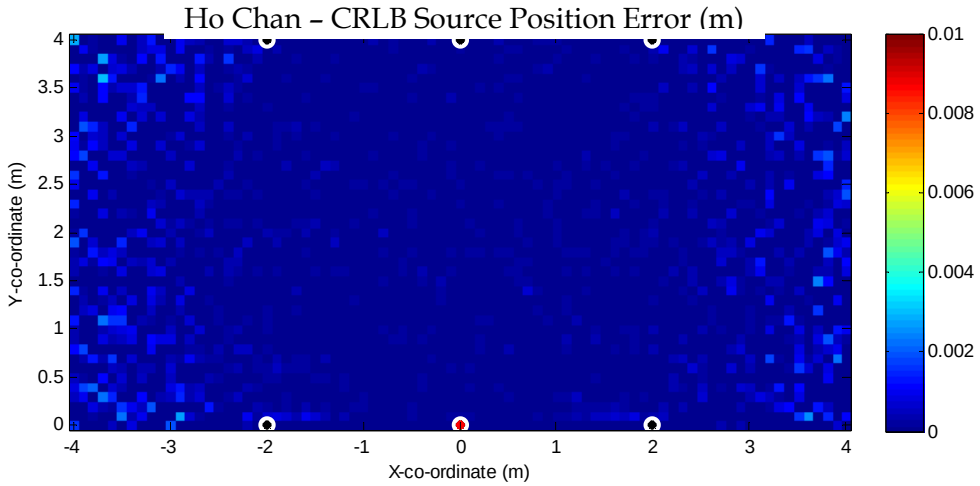

Figure 17. Plot of CRLB subtracted from Ho-Chan standard deviations.

In spite of the superior performance of the Ho-Chan over the SX method, Chan and Ho (1994) proposed that for the case of three sensors the SX method should be used. In simulations the Ho-Chan method does not work with three sensors due to singularities in the calculations. When there are four sensors it was found that the Ho-Chan method had very bad results in areas inside the microphone array as shown in Figure 18. This limits its use to when there are five or more sensors.

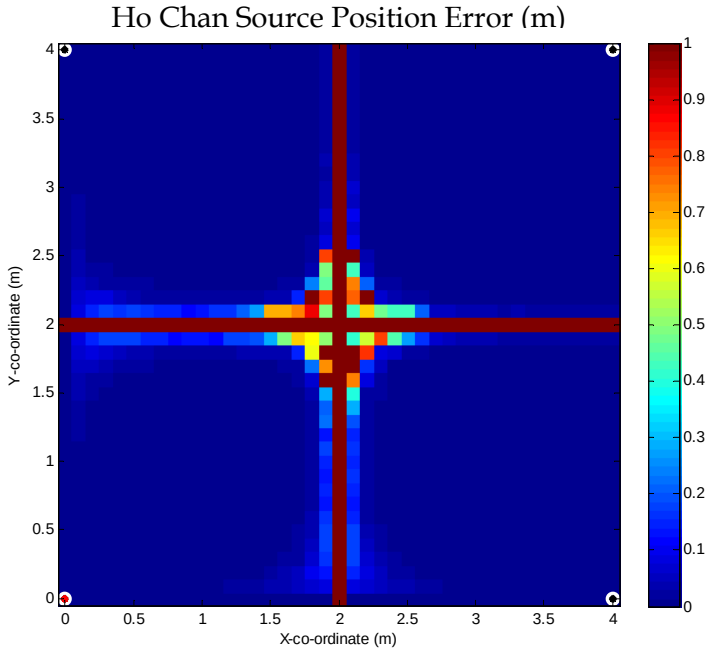

Figure 18. Standard deviation of source position using the Ho-Chan method with four sensors.

4.2 Time-delay Estimators

The Ho-Chan method worked well with error variances close to that of the CRLB provided that the time delays used for calculations are accurate. The main problem is with obtaining these accurate time delays. Room acoustics such as background noise and reverberation are the main contributors to inaccurate time delay estimates. The background noise in the ICS Lab is high due to its computer servers residing in the same room. Reverberation from the walls has been reduced because 60% of the walls have a felt covering. However it still exists due to the uncovered walls and ceiling and from the reflective white-board surfaces (existing along some walls and covering the tables).

The ordinary cross-correlation is a common method used to obtain time-delays between signals (Knapp and Carter 1976). Two alternative time-delay estimation methods have been tested against the ordinary cross-correlation method: the generalised cross-correlation using phase transform (GCC-PHAT) method (Knapp and Carter 1976) and the eigenvalue decomposition (ED) method (Benesty 2000).

4.2.1 Cross-correlation

Cross-correlation of two signals is the convolution of one with the complex conjugate of the other. The cross-correlation $R_{x1,x2}(\mathrm{t})$ of two signals x_1 and x_2 is shown in (15):

$$R_{x1,x2}(\mathrm{t}) = x_1 \bigstar x_2 \equiv x_1(-t) \otimes x_2(t) \tag{15}$$

where $\bigstar$ denotes cross-correlation; $\otimes$ denotes convolution defined by

$$x_1(t) \otimes x_2(t) = \int_{-\infty}^{\infty} x_1(\tau) x_2(t-\tau) d\tau \,. \tag{16}$$

Cross-correlation determines how correlated two signals are. Two signals that are identical but time-shifted by a certain time-delay will have a cross-correlation function with a strong maximum peak at that time-delay point (Stein, 1981). However, when reverberation distorts the input signals, each reverberation causes an extra peak in the cross-correlation function as sidelobes. When the input signal is very periodic the cross-correlation function is also very periodic in form. Each of the periodic peaks would have its own sidelobes which sum with the existing peaks. This makes it difficult to determine which peak is the central time-delay peak and which are just reverberation sidelobes.

Background noise that exists in the signals also affects the plots in two ways depending on whether the noise is random or whether there is a localised noise source. Spatially random noise would create random errors in the plots whereas localised noise would produce extra correlation peaks relative to the time delay of the noise source location. Spatially random noise may arise from a large number of distributed noise sources such as computers; a large number of reflections will also tend to de-correlate the noise sources.

4.2.2 Generalised Cross-correlation

The generalised cross-correlation can be explained using the following equations. The original cross-correlation $R_{x1,x2}(t)$ is related to the cross power spectral density function $G_{x1,x2}(f)$ by the Fourier transform relationship

$$R_{x1,x2}(t) = \int_{-\infty}^{\infty} G_{x1,x2}(f) e^{j2\pi ft} df \ . \tag{17}$$

The generalised cross-correlation has an additional weighting value $\psi(f)$ in the frequency domain, before the integration/summation step and is of the form

$$Rg_{x1,x2}(t) = \int_{-\infty}^{\infty} \psi(f) G_{x1,x2}(f) e^{j2\pi ft} df \ . \tag{18}$$

The weighting value is chosen from a number of different proposed methods, each with their advantages and disadvantages. The ordinary cross-correlation method would have $\psi(f) = 1$. The method chosen for this project is the Phase Transform (GCC-PHAT).

The GCC-PHAT uses a weighting of

$$\psi(f) = \frac{1}{\left|G_{x1,x2}(f)\right|} \tag{19}$$

which produces

$$Rg_{x1,x2}(t) = \int_{-\infty}^{\infty} \frac{G_{x1,x2}(f)}{\left|G_{x1,x2}(f)\right|} e^{j2\pi ft} df \ . \tag{20}$$

Essentially the GCC-PHAT normalises the resulting cross spectral power density of the two signals to a constant value which effectively pre-whitens the cross-correlation function. This pre-whitening equalises the amplitude of the signals across the frequency band leaving only the phase information. This helps to reduce the effects of reverberation on the accuracy of the TDOA estimates (Knapp and Carter 1976).

4.2.3 Eigenvalue Decomposition (ED) Method

The ED method is based on estimating the impulse response between the two signals. A model of the signals received $x_i(n)$, $i = 1, 2$, can be expressed as

$$x_i(n) = g_i \otimes s(n) + b_i(n) \tag{21}$$

where $s_i(n)$ is the source, g_i is the discrete time impulse response of the channel between the source and receiver and $b_i(n)$ is additive noise.

Simplifying (14) by removing additive noise, we have

$$x_1(n) = g_1 \otimes s(n)\,,\; x_2(n) = g_2 \otimes s(n) \tag{22}$$

therefore assuming that the system (room) is linear, time invariant and noise free,

$$x_1(n) \otimes g_2 = g_1 \otimes s(n) \otimes g_2 = g_1 \otimes g_2 \otimes s(n) = g_1 \otimes x_2(n)\,. \tag{23}$$

From this we have the relation

$$\underline{\mathbf{x}}_1^T(n)\mathbf{g}_2 = \underline{\mathbf{x}}_2^T(n)\mathbf{g}_1 \tag{24}$$

where $\underline{\mathbf{x}}_i$ and $\mathbf{g}_i$ are vectors of the signals received and corresponding impulse responses respectively, where $\underline{x}_i(n) = \begin{bmatrix} x_i(n) \\ x_i(n-1) \\ \vdots \\ x_i(n-N+1) \end{bmatrix}$ are the samples along the N tap filter used to model the impulse response.

The covariance matrix of the two signals is

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{x1x1} & \mathbf{R}_{x1x2} \\ \mathbf{R}_{x2x1} & \mathbf{R}_{x2x2} \end{bmatrix} \tag{25}$$

where

$$\mathbf{R}_{xi\,xj} = E\{\mathbf{x}_i(n)\mathbf{x}_j^T(n)\},\; i,j = 1,2. \tag{26}$$

Let

$$\mathbf{u} = \begin{bmatrix} \mathbf{g}_2 \\ -\mathbf{g}_1 \end{bmatrix}. \tag{27}$$

From (24) and (25) we have

$$\begin{aligned} \mathbf{Ru} &= \begin{bmatrix} \mathbf{R}_{x1x1}\mathbf{g}_2 - \mathbf{R}_{x1x2}\mathbf{g}_1 \\ \mathbf{R}_{x2x1}\mathbf{g}_2 - \mathbf{R}_{x2x2}\mathbf{g}_1 \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{E}\{\mathbf{x}_1\mathbf{x}_1^T\}\mathbf{g}_2 - \mathrm{E}\{\mathbf{x}_1\mathbf{x}_2^T\}\mathbf{g}_1 \\ \mathrm{E}\{\mathbf{x}_2\mathbf{x}_1^T\}\mathbf{g}_2 - \mathrm{E}\{\mathbf{x}_2\mathbf{x}_2^T\}\mathbf{g}_1 \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{E}\{\mathbf{x}_1\mathbf{x}_1^T\mathbf{g}_2\} - \mathrm{E}\{\mathbf{x}_1\mathbf{x}_2^T\mathbf{g}_1\} \\ \mathrm{E}\{\mathbf{x}_2\mathbf{x}_1^T\mathbf{g}_2\} - \mathrm{E}\{\mathbf{x}_2\mathbf{x}_2^T\mathbf{g}_1\} \end{bmatrix} \end{aligned}$$

$$= \left[\frac{0}{0}\right] = 0 \tag{28}$$

meaning that vector **u** is the eigenvector of **R** corresponding to an eigenvalue of 0 (Benesty 2000).

u is the estimated impulse response from the source to a microphone. It is calculated using a Least Mean Square (LMS) algorithm with the following equations:

$$e(n) = \mathbf{u}^T(n)\mathbf{x}(n) \tag{29}$$

$$\mathbf{u}(n+1) = \frac{\mathbf{u}(n) - \mu e(n)\mathbf{x}(n)}{\left\|\mathbf{u}(n) - \mu e(n)\mathbf{x}(n)\right\|} \tag{30}$$

where $e(n)$ is the error signal and μ is the step size.

The aim is not to accurately estimate the impulse responses $\mathbf{g}_1$ and $\mathbf{g}_2$ but to find the time-delay. It is sufficient to just detect the two direct paths. By initialising a tap equal to 1 in the middle of the first half of **u** (i.e. in the middle of $\mathbf{g}_2$) and having everything else zero that particular peak will always be dominant. A "mirror" effect will appear in the second half of **u** (i.e. in $\mathbf{g}_1$) in the form of a negative peak that will dominate. The relative position of this peak compared with the original peak in the first half of **u** is the time-delay (Benesty 2000).

4.2.4 Theoretical Performance of Time-delay Estimation

A method for calculating the theoretical performance of time-delay estimation is given by Ardoino, Capriati and Zaccaron (2006) based on cross-correlation. The method assumes an ideal environment with ideal signals. Hence it is useful as a lower bound estimator.

Using the notation in Section 4.2.1 $R_{x1,x1}(t)$ is the cross-correlation of signal x_1 and itself (autocorrelation of x_1). The magnitude of the cross-correlation function $|R_{x1,x1}(t)|$ is approximated as

$$\left|R_{x1,x1}(t)\right| \approx 2P_{x1}\left(1 - \frac{t^2}{2}4\pi^2 {B_{eq}}^2\right) \tag{31}$$

where P_{x1} is the signal power of x_1 and B_{eq} is termed the "Equivalent Bandwidth" given by

$${B_{eq}}^2 = \frac{\int_{-\infty}^{+\infty} f^2 P_S(f)df}{\int_{-\infty}^{+\infty} P_S(f)df}; \tag{32}$$

$P_S(f)$ is the signal Power Spectral Density (PSD) function of x_1 (Ardoino et al. 2006).

Let $B = 1/T_c$ be the sampling frequency where T_c is the sampling time. The signal is valid within the Nyquist frequency band of –B/2 to +B/2 and zero everywhere else.

Assuming that the signal power remains constant (32) now becomes

$$B_{eq}{}^2 = \frac{P_S(f)\cdot\int_{-B/2}^{+B/2} f^2 df}{P_S(f)\cdot\int_{-B/2}^{+B/2} 1\ df} = \frac{\left[\frac{1}{3}f^3\right]_{-B/2}^{+B/2}}{[f]_{-B/2}^{+B/2}} = \frac{\frac{1}{3}\left(\frac{B^3}{8}+\frac{B^3}{8}\right)}{\frac{B}{2}+\frac{B}{2}} = \frac{B^2}{12}. \quad (33)$$

The analysis of the cross-correlation function is based on taking three points of the function at time lags $-MT_c$, 0 and $+MT_c$ where M is a chosen number of samples either side of the maximum peak. The peak is assumed to always lie in the time-span of $2MT_c$. A parabola is fitted to these three points and the maximum of the parabola corresponds to the time-delay estimate (Ardoino et al. 2006).

The standard deviation of the time-delay estimates σ_{TDOA} is given by

$$\sigma_{TDOA} \approx \frac{1}{\sqrt{N}\cdot 2\pi\cdot B_{eq}}\cdot\sqrt{\frac{1}{SNR^2}\left(\frac{1}{8\pi^2\cdot B_{eq}{}^2\cdot M^2 T_c^2}\right)+\frac{2}{SNR}} \quad (34)$$

where N is the length of the signal in number of samples and SNR is the Signal-to-Noise Ratio.

Given (33), (34) can be simplified as

$$\sigma_{TDOA} \approx \frac{\sqrt{12}}{\sqrt{N}\cdot 2\pi\cdot B}\cdot\sqrt{\frac{1}{SNR^2}\left(\frac{12}{8\pi^2\cdot M^2}\right)+\frac{2}{SNR}}. \quad (35)$$

σ_{TDOA} gives a standard deviation of time in seconds. Setting B to 1 will give a standard deviation in number of samples.

The standard deviation was calculated for each of the SNR's and integration periods used in the experiments in section 6 of this report. It was found that in each case the standard

deviation was less than 0.2 of a sample. The actual standard deviations are shown in Table 11 in section 6. Thus uncorrelated noise on the microphones is not expected to limit the accuracy of the timing estimate in the experiments carried out in this report. This was not observed in practice. Reasons for this discrepancy include reverberation and the fact that the noise in the room is coming mainly from a localised source and hence is not independent on each microphone, as assumed by the model.

5. Implementation

The localisation system was implemented in three parts: the localisation algorithms which produces the source position estimate based on the relative time-delays between receivers, the time-delay estimators which produce the relative time-delays for the localisation algorithms, and an iterator to repeat the location estimation for tracking purposes.

5.1 Localisation Algorithms and Iterator

The SX and Ho-Chan methods were implemented in MATLAB. This system was designed to be used in the ICS Laboratory using four microphones placed along the walls so only the SX method was used. It takes inputs of the four microphone coordinates and the relative time-delays produced by the time-delay estimators described in Section 5.2. Once the calculation is complete the algorithm chooses the coordinates based on the roots of the equation as described in Section 4.1.1.

The iterator runs a time-delay estimator for each receiver pair and the localisation algorithm multiple times to track the location of the source. Two iterators were created, one which reads four Microsoft Wave computer sound files as inputs for testing purposes, the other to read direct from the ASIO interface of the audio I/O card on the computer. The localisation algorithm contains only small matrix operations which amount to less than 100 multiplications per iteration. Most of the computation is done for the time-delay estimators which process larger amounts of information as described in Section 5.2.

The iterator reads directly from the computer I/O card using a set of MATLAB 'mex' code files called pa_wavplay (http://sourceforge.net/projects/pa-wavplay/). pa_wavplay was developed by Johnson Chen and Matt Frear based on a free open-source C library called PortAudio (http://www.portaudio.com/). The file set includes functions to record and playback audio using the available I/O interfaces on the computer.

5.2 Time-delay Estimators

Both the GCC-PHAT and ED methods were implemented in MATLAB using Fast Fourier Transforms (FFTs). Both methods take in two arrays of data and the relative time-delay between them is found. They are used similarly to the 'xcorr' (cross-correlation) function of MATLAB.

The ED method includes an LMS algorithm that was based on the Block LMS (Fast LMS) algorithm from Haykin (1991). In addition to the algorithm itself a simple pre-whitening process (similar to that in GCC-PHAT) was implemented on the data that was input into algorithm.

GCC-PHAT is much faster since it only requires 12L multiplication operations per estimate while ED requires (floor(L/N)-1)(38N)+12N multiplications, where L is length of the input signals, N is length of vector **u**, "floor" indicates rounding down to the nearest integer. For 1 second worth of data at 44100Hz sampling rate and length of **u** set at 1580,

GCC-PHAT requires 529200 multiplications while ED requires 1580000 multiplications. For 0.5 second data this becomes 264600 and 739440 respectively.

A rough estimate of ED calculations by assuming N is fixed and removing the -1 means it has about 38L+12N operations, or roughly 3 times the number of operations in GCC-PHAT since N is much less than L in most cases. Hence based on computational requirements GCC-PHAT would be chosen over the ED method. This may not be a problem in modern computers if the real-time sampling and calculations were conducted in parallel, however if conducted in series then a longer calculation time would reduce the number of available estimates in a fixed time frame.

The ED method is regraded to be more robust to reverberation than the GCC-PHAT, since instead of finding the maximum peak (which is how correlation-based methods work), it tries to estimate the impulse response of the room which includes the time-delay peak as well as reverberation peaks. From this estimate it finds the direct path peak by choosing the first main peak in the impulse response.

6. Experiments

Sound recordings were done in the ICS Laboratory using the four existing Shure Microflex MX393/O omnidirectional microphones in the room and a Behringer ADA8000 8-channel A-D/D-A converter. The microphones were placed on the walls of the room in a configuration shown in Figure 19 at a height of 2.3 metres. The samples were recorded straight into a PC computer by connecting the A-D converter to a RME HDSP 9652 audio I/O card via fibre optic cable using ASIO interface.

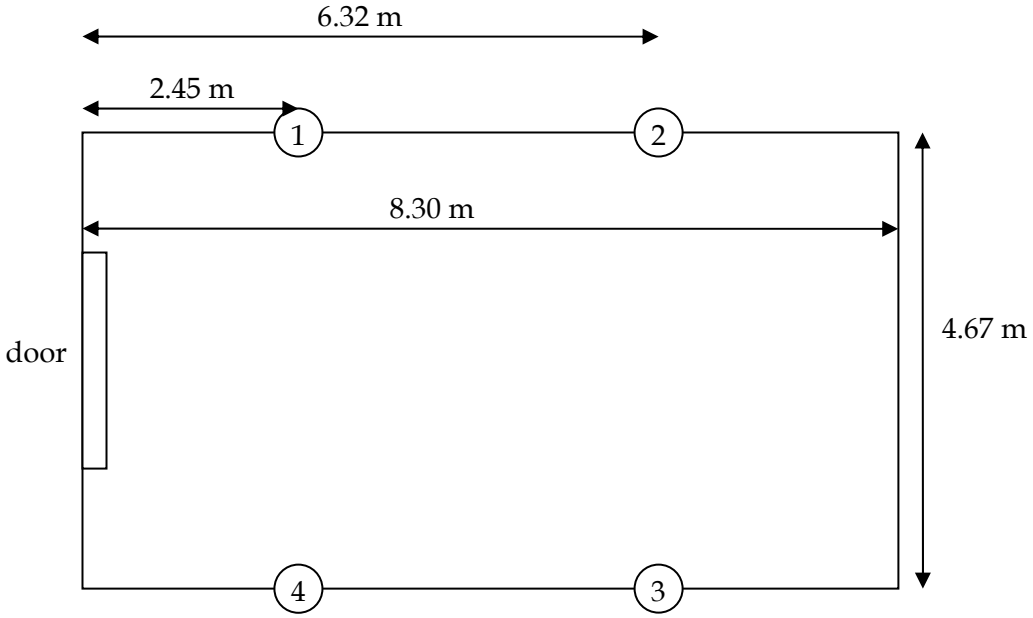

Figure 19. Microphone configuration in ICS Laboratory.

The recordings were of a Gaussian random white noise source generated from MATLAB and played through a Sony SRS-PC300D speaker with a frequency response of 50-20000Hz. Different sound levels were recorded. The noise source was then filtered to different frequency band outputs and this was also recorded. The output and input sampling rates were both 44100Hz.

Table 1 shows all the recordings and their Signal-to-Noise Ratios (SNRs). The sound level was detected using a Castle GA208 Sound Level Meter held 2cm from the sound source (speaker). "Overload" represents when the sound meter was saturated. It means the sound level was too high for the meter to detect properly, and for most cases is around 100dBA.

Table 1. Samples recorded from the ICS Laboratory.

Frequency (Hz)	Sound levels (dBA)	SNRs (dB)
White (50-20000)	Overload	10.05
	90	-0.3309
	80	-10.33
	70	-20.33
50-500	95	10.37
	90	8.110
	80	-1.890
	70	-11.89
500-2000	Overload	7.736
	90	-5.698
	80	-15.70
	70	-25.70
2000-5000	Overload	9.044
	90	-7.494
	80	-17.49
	70	-27.49
5000-10000	Overload	12.75
	90	2.969
	80	-7.031
	70	-17.03
10000-15000	Overload	13.28
	90	8.169
	80	-1.831
	70	-11.83
15000-20000	75	0.7958
	70	-4.204

7. Results

Table 2 shows the true time-delays for each microphone pair in samples.

Table 2. True time-delays for each microphone pair in samples (sampling rate 44100 Hz).

Mic 1	Mic 2	Time-delay (samples)
1	2	13
2	3	119
3	4	-9
1	4	123
1	3	132
2	4	110

Table 3 and Table 4 show the estimated time-delays from the three estimation methods. The tables show that GCC-PHAT had the most accurate results out of the three by the largest number of estimates close to the true time-delays. It was surprising to see that none of the time-delay estimators worked well in the frequency band 50-2000Hz.

Frequency spectra of the received signals were created to determine the cause of the problems in the 50-2000Hz band. These spectra plots are shown in Figures 66-91 in Appendix C. The most notable observation from the plots is the high background noise that exists in the 0-1500Hz area as shown in Figure 20. This is the most likely reason that even GCC-PHAT was unable to produce accurate time-delays in this frequency band.

Table 3. Time-delays between pairs of the four microphones estimated using three methods: cross-correlation (xcorr), GCC-PHAT (gcorr) and ED (eigen). "max length" indicates the maximum length of the sample in seconds. All data used were sampled at 44100 Hz. "n" is the length of data used for estimating the time delays in samples. The time-delays are expressed in samples. The shaded time-delay estimates are within 5 samples of the true time-delays.

white

xcorr

white	overload	90dBA	80dBA	70dBA
max length	15	14	15	15
n	651050	605950	651050	649050
x12	13	964	996	976
x23	119	119	120	-43150
x34	-9	-9	618	51121
x14	123	9050	9061	-9092
x13	132	-736	36825	584
x24	110	657	618	604

gcorr

white	overload	90dBA	80dBA	70dBA
max length	15	14	15	15
n	651050	605950	651050	649050
g12	13	13	13	13
g23	119	119	120	119
g34	-10	-10	-10	-10
g14	122	123	123	123
g13	132	133	133	133
g24	110	110	110	110

eigen

white	overload	90dBA	80dBA	70dBA
max length	15	14	15	15
n	651050	605950	651050	649050
e12	13	13	14	13
e23	119	120	120	119
e34	-10	-10	-10	-10
e14	122	123	123	123
e13	132	133	133	133
e24	110	110	26	110

50-500Hz

xcorr

50-500Hz	95dBA	90dBA	80dBA	70dBA
max length	14	15	15	14
n	606950	500000	651050	606950
x12	-1	-3	-2532	-2535
x23	33	33	33	25172
x34	-11	-11	-11	-4364
x14	19	19	-457	5416
x13	-267	-269	602	-733
x24	1174	1173	666	670

gcorr

50-500Hz	95dBA	90dBA	80dBA	70dBA
max length	14	15	15	14
n	606950	500000	651050	606950
g12	0	0	-4	191
g23	128	0	14	14
g34	0	-11	-10	473
g14	1	366	166	286
g13	1544	0	-144	-1045
g24	1544	484	599	600

eigen

50-500Hz	95dBA	90dBA	80dBA	70dBA
max length	14	15	15	14
n	606950	500000	651050	606950
e12	0	165	-345	542
e23	0	-45	-240	262
e34	0	461	303	451
e14	-400	-352	29	411
e13	0	428	-151	37
e24	-400	500	375	-190

500-2000Hz

xcorr

500-2000Hz	overload	90dBA	80dBA	70dBA
max length	14	15	15	14
n	606950	650050	651050	606950
x12	-1051	-2503	-2507	-2499
x23	124	25166	98	25153
x34	-11	-11	482	-1008
x14	-36	9049	9061	9068
x13	133	-686	8096	36762
x24	490	655	654	676

gcorr

500-2000Hz	overload	90dBA	80dBA	70dBA
max length	14	15	15	14
n	606950	650050	651050	606950
g12	0	-352	198	180
g23	350	350	351	-339
g34	-9	-9	882	-11
g14	126	126	1104	616
g13	474	383	1	-1026
g24	0	730	730	729

eigen

500-2000Hz	overload	90dBA	80dBA	70dBA
max length	14	15	15	14
n	606950	650050	651050	606950
e12	446	-188	289	-184
e23	154	-113	377	-131
e34	364	473	78	-263
e14	286	-506	518	-152
e13	-410	109	310	-526
e24	466	-401	-446	-254

2000-5000Hz

xcorr

2000-5000Hz	overload	90dBA	80dBA	70dBA
max length	15	15	15	15
n	651050	650050	651050	651050
x12	13	1022	31465	-2525
x23	119	119	118	25192
x34	-10	-11757	-4360	-4347
x14	123	-53931	-9031	9055
x13	133	629	610	-737
x24	109	654	637	657

gcorr

2000-5000Hz	overload	90dBA	80dBA	70dBA
max length	15	15	15	15
n	651050	650050	651050	651050
g12	13	13	13	-45
g23	119	119	119	119
g34	-10	-10	-10	-9
g14	123	123	123	123
g13	132	132	132	132
g24	110	110	110	109

eigen

2000-5000Hz	overload	90dBA	80dBA	70dBA
max length	15	15	15	15
n	651050	650050	651050	651050
e12	-97	-118	-45	-47
e23	120	520	-541	117
e34	-9	-374	120	70
e14	123	123	124	-410
e13	-384	320	-545	133
e24	191	108	-119	68

Table 4. Estimated time-delays using three methods: cross-correlation (xcorr), generalised cross-correlation (gcorr) and eigenvalue decomposition (eigen). Continued from Table 3.

5000-10000Hz

xcorr

5000-10000Hz	overload	90dBA	80dBA	70dBA
max length	15	15	14	15
n	650050	591050	606950	651050
x12	13	13	997	1004
x23	114	114	114	98533
x34	-4	-9	-45831	-22431
x14	123	123	9050	16186
x13	132	132	-735	656
x24	110	110	36789	662

gcorr

5000-10000Hz	overload	90dBA	80dBA	70dBA
max length	15	15	14	15
n	650050	591050	606950	651050
g12	13	13	13	13
g23	119	114	114	114
g34	-10	-9	-9	-9
g14	123	123	123	123
g13	132	132	132	132
g24	110	110	110	110

eigen

5000-10000Hz	overload	90dBA	80dBA	70dBA
max length	15	15	14	15
n	650050	591050	606950	651050
e12	13	13	-45	14
e23	119	119	115	114
e34	-10	-9	-9	-9
e14	123	123	129	329
e13	132	127	64	132
e24	110	110	110	110

10000-15000Hz

xcorr

10000-15000Hz	94dBA	90dBA	80dBA	70dBA
max length	14	15	15	14
n	606950	580000	651050	606950
x12	13	9	13	-2509
x23	123	123	123	84
x34	-13	-13	-13	-4358
x14	4	126	-546	9076
x13	136	132	133	18713
x24	-59	110	635	51242

gcorr

10000-15000Hz	94dBA	90dBA	80dBA	70dBA
max length	14	15	15	14
n	606950	580000	651050	606950
g12	13	13	13	13
g23	119	119	123	123
g34	-10	-10	-13	-13
g14	123	184	123	123
g13	132	132	133	194
g24	110	110	110	110

eigen

10000-15000Hz	94dBA	90dBA	80dBA	70dBA
max length	14	15	15	14
n	606950	580000	651050	606950
e12	13	461	451	13
e23	119	119	123	123
e34	-10	-10	-13	-13
e14	123	148	519	123
e13	133	132	133	129
e24	110	110	110	-66

15000-20000Hz

xcorr

15000-20000Hz	75dBA	70dBA		
max length	14	14		
n	300000	450000		
x12	947	951		
x23	119	119		
x34	-7	-7		
x14	125	125		
x13	118	118		
x24	707	642		

gcorr

15000-20000Hz	75dBA	70dBA		
max length	14	14		
n	300000	450000		
g12	-48	-48		
g23	119	119		
g34	-7	-7		
g14	122	122		
g13	132	132		
g24	107	107		

eigen

15000-20000Hz	75dBA	70dBA		
max length	14	14		
n	300000	450000		
e12	9	13		
e23	119	119		
e34	-7	-9		
e14	184	122		
e13	132	132		
e24	109	107		

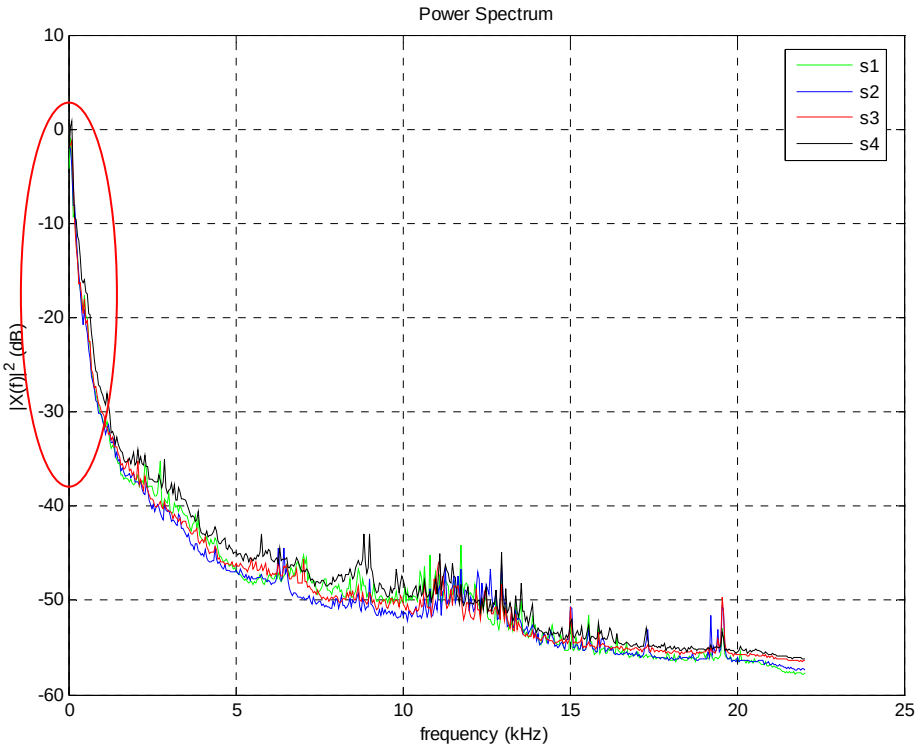

Figure 20. Frequency spectrum of recorded background noise in the ICS Laboratory. The red oval indicates the dominant low frequencies.

The conclusion from the analysis so far is that the GCC-PHAT is the best of the three time-delay estimators. However this analysis uses large amounts of data. In real-time localisation the system can only collect small amounts of data at a time so that the location estimate can be constantly and quickly updated.

The next analysis tested the estimators' abilities to work with small amounts of data. The same data were broken into small parts and given to the estimators to find the time-delay of each part. The white noise samples of all four sound levels (overload, 90dBA, 80dBA and 70dBA) were used for this.

Initially the 15 second samples were broken into 1 second samples. The time-delay estimates were found from each sample and plotted. Figures 92-95 in Appendix D show that while all three estimators worked well for 'overload' white noise, when the sound level is reduced the cross-correlation performs badly. In contrast both GCC-PHAT and ED methods performed very well having at most only 2 outliers in a plot. The samples were then broken into 0.1 second samples mainly to find whether GCC-PHAT or ED method is better when less information is available. Cross-correlation was also tested as a reference. These plots are shown in Figures 96-99 in Appendix D.

Table 5 and Table 6 summarise the number of outliers in each set of estimates.

Table 5. Number of outliers for estimates of 1 second data length for cross-correlation (xcorr), GCC-PHAT (gcorr) and ED method (eigen) shown for each microphone pair.

Estimates with 1 second length data, 14 estimates per set								
	Microphone pairs →	1 2	2 3	3 4	1 4	1 3	2 4	total
xcorr	overload	0	0	0	0	0	0	0
	90 dBA	12	0	1	14	11	14	52
	80 dBA	14	5	11	14	14	14	72
	70 dBA	14	8	14	14	14	14	78
gcorr	overload	0	0	0	0	0	0	0
	90 dBA	0	0	0	0	0	0	0
	80 dBA	1	0	0	0	0	0	1
	70 dBA	0	0	0	0	0	0	0
eigen	overload	0	0	0	0	0	0	0
	90 dBA	0	0	0	0	0	0	0
	80 dBA	2	0	0	0	0	0	2
	70 dBA	0	0	0	0	0	0	0

Table 6. Number of outliers for estimates of 0.1 second data length for cross-correlation (xcorr), GCC-PHAT (gcorr) and ED method (eigen) shown for each microphone pair.

Estimates with 0.1 second length data, 147 estimates per set								
	Microphone pairs →	1 2	2 3	3 4	1 4	1 3	2 4	total
xcorr	overload	0	0	0	0	0	0	0
	90 dBA	124	66	44	121	105	141	601
	80 dBA	140	104	128	146	143	147	808
	70 dBA	147	136	143	147	147	147	867
gcorr	overload	0	0	0	0	0	0	0
	90 dBA	1	0	0	0	0	0	1
	80 dBA	4	0	0	0	0	0	4
	70 dBA	6	10	0	8	9	0	33
eigen	overload	0	0	0	0	0	0	0
	90 dBA	12	0	0	0	1	0	13
	80 dBA	17	0	0	0	3	0	20
	70 dBA	10	7	0	1	16	0	34

Table 5 and Table 6 clearly show that the ordinary cross-correlation works only for 'overload' sound levels, for lower signal power it is unable to provide correct estimates for the majority of the time. GCC-PHAT and ED methods worked much better and only to show small numbers of outliers in the 70 dBA and 80 dBA levels. Overall GCC-PHAT has less outliers than the ED method.

One observation from the plots in Figures 92-96 is that there is a repeating striation effect in the cross-correlation plots, more notably in Figures 97-99. This effect is probably caused

by correlations in the background noise in the room, but could also be due to reverberation. This effect is reduced in GCC-PHAT and ED plots possibly due to the pre-whitening step, which effectively flattens the background noise spectrum, thus reducing the relative background noise power. The pre-whitening filter also reduces the side-lobe levels in the auto-correlation function of both the desired and background noise signals.

Tables 7-10 show some statistical information obtained from the analysis and includes the mean, median and variance of estimates. The column headings 'diffmean' and 'diffmedian' show the mean and median after the true values have been subtracted so that deviations can be seen without any bias.

Tables 7 and 8 show information for the time-delay estimates of 1 second and 0.1 second length data respectively.

Table 7. Statistical information for time-delay estimates of 1 second length data. Values are shown in samples, sample rate 44100Hz. 'x' represents cross-correlation, 'g' represents GCC-PHAT and 'e' represents ED method. The numbers next to the letters indicate microphone pairs, e.g. x12 means the estimates for microphone pair 1 and 2 using cross-correlation.

white overload

	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x12	13	13	0	0	0	g12	13	13	0	0	0	e12	13	13	0	0	0
x23	119	119	0	0	0	g23	119	119	0	0	0	e23	119	119	0	0	0
x34	-9	-9	0	0	0	g34	-9.5	-9.5	0.2692	-0.5	-0.5	e34	-9.5714	-10	0.2637	-0.5714	-1
x14	122.8571	123	0.1319	-0.1429	0	g14	122.4298	122	0.2637	-0.5702	-1	e14	122.4286	122	0.2637	-0.5714	-1
x13	132	132	0	0	0	g13	132	132	0	0	0	e13	132	132	0	0	0
x24	110	110	0	0	0	g24	110	110	0	0	0	e24	109.8571	110	0.1319	-0.1429	0

white 90dBA

	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x12	345.1429	964	1553300	332.1429	951	g12	13.2143	13	0.1813	0.2143	0	e12	13.2143	13	0.1813	0.2143	0
x23	119.1429	119	0.1319	0.1429	0	g23	119.1429	119	0.1319	0.1429	0	e23	119.0714	119	0.0714	0.0714	0
x34	-82.1429	-9	74741	-73.1429	0	g34	-9.9286	-10	0.0714	-0.9286	-1	e34	-9.7857	-10	0.1813	-0.7857	-1
x14	-155.2143	-481	2907000	-278.2143	-604	g14	123.0714	123	0.0714	0.0714	0	e14	123.0714	123	0.0714	0.0714	0
x13	-689.2143	-706	516810	-821.2143	-838	g13	132.9286	133	0.0714	0.9286	1	e13	132.9286	133	0.0714	0.9286	1
x24	634	654.5	2174.3	524	544.5	g24	110	110	0	0	0	e24	110	110	0	0	0

white 80dBA

	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x12	239.2857	996	2241800	226.2857	983	g12	9.1429	13	243.0549	-3.8571	0	e12	42.0714	13	15697	29.0714	0
x23	-316.8571	120	3747900	-435.8571	1	g23	119.7857	120	0.1813	0.7857	1	e23	119.7143	120	0.2198	0.7143	1
x34	-106.0714	-9	11844000	-97.0714	0	g34	-10	-10	0	-1	-1	e34	-10	-10	0	-1	-1
x14	-1533.9	-1908.5	26764000	-1656.9	-2031.5	g14	123	123	0	0	0	e14	123	123	0	0	0
x13	-1305.6	-1802.5	6269500	-1437.6	-1934.5	g13	133	133	0	1	1	e13	133	133	0	1	1
x24	876.5714	618.5	1352600	766.5714	508.5	g24	110	110	0	0	0	e24	110	110	0	0	0

white 70dBA

	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x12	725.4286	974.5	884640	712.4286	961.5	g12	13.0714	13	0.0714	0.0714	0	e12	13.0714	13	0.0714	0.0714	0
x23	226.9286	119.5	4980000	107.9286	0.5	g23	119.1429	119	0.1319	0.1429	0	e23	119.1429	119	0.1319	0.1429	0
x34	-1994.6	-1014.5	12877000	-1985.6	-1005.5	g34	-10	-10	0	-1	-1	e34	-10	-10	0	-1	-1
x14	-594.7857	-1196	14322000	-717.7857	-1319	g14	123	123	0	0	0	e14	123	123	0	0	0
x13	-904.6429	-711.5	5234200	-1036.643	-843.5	g13	132.7857	133	0.1813	0.7857	1	e13	132.8571	133	0.1319	0.8571	1

Table 8. Statistical information for time-delay estimates of 0.1 second length data. Values are shown in samples, sample rate 44100Hz. 'x' represents cross-correlation, 'g' represents GCC-PHAT and 'e' represents ED method. The numbers next to the letters indicate microphone pairs, e.g. x12 means the estimates for microphone pair 1 and 2 using cross-correlation.

white overload

	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x12	13	13	0	0	0	g12	13	13	0	0	0	e12	13	13	0	0	0
x23	119	119	0	0	0	g23	119	119	0	0	0	e23	119	119	0	0	0
x34	-9.0068	-9	0.0068	-0.0068	0	g34	-9.4422	-9	0.2483	-0.4422	0	e34	-9.4966	-9	0.2517	-0.4966	0
x14	122.8367	123	0.1375	-0.1633	0	g14	122.3605	122	0.2321	-0.6395	-1	e14	122.3469	122	0.2281	-0.6531	-1
x13	132	132	0	0	0	g13	132	132	0	0	0	e13	132	132	0	0	0
x24	110	110	0	0	0	g24	109.9388	110	0.0579	-0.0612	0	e24	109.9184	110	0.0755	-0.0816	0

white 90dBA

	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x12	197.1361	98	875600	184.1361	85	g12	12.7891	13	23.1813	-0.2109	0	e12	8.4558	13	255.8936	-4.5442	0
x23	81.7415	119	103480	-37.2585	0	g23	119.2177	119	0.1715	0.2177	0	e23	119.2653	119	0.1963	0.2653	0
x34	-144.2653	-9	343640	-135.2653	0	g34	-9.9048	-10	0.0868	-0.9048	-1	e34	-9.9116	-10	0.0812	-0.9116	-1
x14	107.2653	123	484500	-15.7347	0	g14	123.0476	123	0.0457	0.0476	0	e14	123.0476	123	0.0457	0.0476	0
x13	-160.0952	132	391600	-292.0952	0	g13	132.8503	133	0.1281	0.8503	1	e13	132.3333	133	42.4292	0.3333	1
x24	56.9728	110	700430	-53.0272	0	g24	110	110	0	0	0	e24	110	110	0	0	0

white 80dBA

	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x12	204.3946	569	1566800	191.3946	556	g12	11.6599	13	90.5958	-1.3401	0	e12	6.5374	13	349.8941	-6.4626	0
x23	93.3878	119	323200	-25.6122	0	g23	119.7687	120	0.179	0.7687	1	e23	119.7619	120	0.1826	0.7619	1
x34	-41.3061	-9	935590	-32.3061	0	g34	-10	-10	0	-1	-1	e34	-10	-10	0	-1	-1
x14	-40.6735	-142	1357400	-163.6735	-265	g14	123	123	0	0	0	e14	123.0068	123	0.0068	0.0068	0
x13	37.1973	562	1101000	-94.8027	430	g13	132.9456	133	0.0518	0.9456	1	e13	131.3061	133	125.433	-0.6939	1
x24	-19.3741	206	1039200	-129.3741	96	g24	110	110	0	0	0	e24	110	110	0	0	0

white 70dBA

	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x12	-0.7007	-24	1747300	-13.7007	-37	g12	12.4966	13	6.7175	-0.5034	0	e12	5.9932	13	1946.6	-7.0068	0
x23	128.8776	119	450710	9.8776	0	g23	141.1224	119	126380	22.1224	0	e23	115.7483	119	1764.6	-3.2517	0
x34	106.4286	36	1322400	115.4286	45	g34	-9.9796	-10	0.0201	-0.9796	-1	e34	-9.9524	-10	0.0457	-0.9524	-1
x14	-131.8912	-85	853120	-254.8912	-208	g14	116.3265	123	784.1666	-6.6735	0	e14	123.0884	123	0.6976	0.0884	0
x13	-1.0544	87	1286800	-133.0544	-45	g13	124.6803	133	1020.9	-7.3197	1	e13	129.0884	133	3640.4	-2.9116	1
x24	55.8639	505	927260	-54.1361	395	g24	110	110	0	0	0	e24	110	110	0	0	0

The true coordinates of the source based on the true time-delays is (4.1579, 1.7109). Localisation estimates were done from these time-delay estimates using the SX method. Table 9 and Table 10 show statistical information for the localisation estimates using the time-delay estimates for 1 second data and 0.1 second data respectively. 'diffmean' and 'diffmedian' represent the same as that for Tables 7 and 8 where the true values were subtracted from the mean and median to remove bias.

Table 9. Statistical information for localisation coordinate estimates using time-delay estimates of 1 second length data. Coordinates x and y are shown in metres. 'xcorr', 'gcorr' and 'eigen' represent cross-correlation, GCC-PHAT and ED methods respectively.

xcorr						gcorr						eigen					
white overload																	
	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x	4.1581	4.1579	3.95E-07	0.0002	0	x	4.1589	4.1596	7.91E-07	0.001	0.0017	x	4.1589	4.1596	7.91E-07	0.001	0.0017
y	1.7114	1.7109	1.41E-06	0.0005	0	y	1.7128	1.7142	2.82E-06	0.0019	0.0033	y	1.7128	1.7142	2.82E-06	0.0019	0.0033
white 90dBA																	
	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x	12.5613	0.4353	3724.7	8.4034	-3.7226	x	4.1604	4.1597	3.78E-06	0.0025	0.0018	x	4.1604	4.1597	3.78E-06	0.0025	0.0018
y	30.9001	3.9712	9334.9	29.1892	2.2603	y	1.7092	1.709	8.45E-07	-0.0017	-0.0019	y	1.7092	1.709	8.45E-07	-0.0017	-0.0019
white 80dBA																	
	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x	18.018	-13.7524	9704.7	13.8601	-17.9103	x	4.1405	4.1597	0.006	-0.0174	0.0018	x	4.2499	4.1597	0.1782	0.092	0.0018
y	90.1342	34.3648	20002	88.4233	32.6539	y	1.7019	1.709	8.00E-04	-0.009	-0.0019	y	1.7355	1.709	0.017	0.0246	-0.0019
white 70dBA																	
	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x	10.343	2.8973	4365.6	6.1851	-1.2606	x	4.1596	4.1597	2.46E-06	0.0017	0.0018	x	4.1598	4.1597	2.21E-06	0.0019	0.0018
y	51.4181	20.6387	9837.9	49.7072	18.9278	y	1.7095	1.709	7.83E-07	-0.0014	-0.0019	y	1.7094	1.709	6.31E-07	-0.0015	-0.0019

Table 10. Statistical information for localisation coordinate estimates using time-delay estimates of 0.1 second length data. Coordinates x and y are shown in metres. 'xcorr', 'gcorr' and 'eigen' represent cross-correlation, GCC-PHAT and ED methods respectively.

xcorr						gcorr						eigen					
white overload																	
	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x	4.1582	4.1579	4.1242E-07	0.0003	0	x	4.159	4.1596	6.96E-07	0.0011	0.0017	x	4.159	4.1596	6.84E-07	0.0011	0.0017
y	1.7115	1.7109	1.4712E-06	0.0006	0	y	1.713	1.7142	2.48E-06	0.0021	0.0033	y	1.7131	1.7142	2.44E-06	0.0022	0.0033
white 90dBA																	
	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x	0.7817	2.4641	40.7387	-3.3762	-1.6938	x	4.1582	4.1597	5.70E-04	0.0003	0.0018	x	4.1358	4.1597	0.0064	-0.0221	0.0018
y	2.8046	2.8069	16.7858	1.0937	1.096	y	1.7087	1.709	7.69E-05	-0.0022	-0.0019	y	1.7019	1.709	0.001	-0.009	-0.0019
white 80dBA																	
	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x	-1.9362	1.3293	293.0539	-6.0941	-2.8286	x	4.1529	4.1597	0.0022	-0.005	0.0018	x	4.1244	4.1597	0.0089	-0.0335	0.0018
y	5.1116	4.0725	170.3808	3.4007	2.3616	y	1.7066	1.709	0.000299	-0.0043	-0.0019	y	1.7006	1.709	0.0017	-0.0103	-0.0019
white 70dBA																	
	mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian		mean	median	variance	diffmean	diffmedian
x	-2.437	0.4154	249.4983	-6.5949	-3.7425	x	4.1544	4.1597	0.0073	-0.0035	0.0018	x	4.0795	4.1597	0.3874	-0.0784	0.0018
y	4.6517	3.6709	77.3434	2.9408	1.96	y	1.7491	1.709	0.0135	0.0382	-0.0019	y	1.6864	1.709	0.1324	-0.0245	-0.0019

Tables 7 and 8 show that the ordinary cross-correlation has mean and median values with high deviation from the true values and very high variance for all data other than that of 'overload' noise level. GCC-PHAT and ED methods held up well with medians differing by 1 sample at the most from the true value; however the outliers created mean and variance values that detracted away from their overall effectiveness shown in the plots in Figures 92-99. The coordinate estimates in Tables 9 and 10 reflect what was shown in Tables 7 and 8 with GCC-PHAT and ED methods producing good estimates the majority of the time while the cross-correlation only does so for 'overload' sound level data. Overall GCC-PHAT had the best statistical results.

The theoretical standard deviations were calculated for 1 second and 0.1 second estimates using equation (35) and are shown in Table 11 in units of samples. Note that the theoretical standard deviation is significantly less than that measured in practice, for example for the 70 dBA power it is 0.17 samples^2 which corresponds to 7.6E-6 m^2 at a 44 KHz sample rate, which is significantly less than 0.013 m^2 in Table 10. Overall the theoretical standard deviation remains much less than one sample and hence should not affect the accuracy of the measurement. The increase in variance in the practical system is thought to be due to reverberation and the fact that the background noise was generated by a localised sound source and hence was not completely un-correlated between microphones.

Table 11. Theoretical standard deviations for 1 second and 0.1 second TDOA estimates in samples.

Sound level of sample	Standard deviations (samples) 1 sec	Standard deviations (samples) 0.1 sec
Overload	1.1675e-003	3.6919e-003
90 dBA	3.8746e-003	1.2253e-002
80 dBA	1.2741e-002	4.0289e-002
70 dBA	5.3322e-002	0.1686

8. Conclusions

Two speaker localisation algorithms, the Spherical Intersection (SX) method and the Ho-Chan method, were implemented in MATLAB. Simulations were done to test their effectiveness to locate a source in a virtual room. Both methods were successful in locating a source anywhere in the room provided that the time-delay estimates were accurate, although the Ho-Chan method has smaller error variances.

Three time-delay estimators were compared: the ordinary cross-correlation, the generalised cross-correlation using phase transform (GCC-PHAT), and eigenvalue decomposition (ED) method. GCC-PHAT and ED were implemented in MATLAB. From the experiments with the three time-delay estimators, it is clear that the ordinary cross-correlation is not able to work well in environments like the ICS Laboratory where there is reverberation and significant background noise. The GCC-PHAT method, although only an extension to the ordinary cross-correlation, works well in the face of background noise and is good for real-time applications due to its low calculation time. The ED method is a strong competitor to the GCC-PHAT due to its similar performance. However it has a higher computation requirement and would be slower than GCC-PHAT. Hence GCC-PHAT would be better suited for real-time use.

9. Future Developments

The system in its current state is incomplete. The objective of the research is to eventually have a system that will detect the location of a person speaking in real-time. There are many areas that can be improved with further research and development.

The Ho-Chan method is good provided that the time-delays are accurate. Currently the algorithm is implemented for two-dimensional localisation. This can produce errors since the speaker sources and the microphone receivers do not always exist in the same 2D plane due to height differences (such as the difference between a sitting speaker and a standing speaker). An implementation for three dimensions would therefore provide more accurate estimates.

The ED method was originally designed to be an estimator that is robust to reverberation (Benesty 2000). ED is a relatively new method compared with GCC-PHAT and hence it is likely to be improved with further development. Research can be done to improve its robustness to both background noise and reverberation while being able to work with smaller amounts of data to increase speed.

Further improvement of time-delay estimation can be accomplished using noise reduction or noise cancelling techniques. These can remove or lessen much of the background noise that has an adverse effect on the time-delay estimators.

A Kalman filter tries to minimise the mean squared error to estimate the state of a process, and supports estimation of past, present and future states (Welch and Bishop 1995). The implementation of a Kalman filter in the system should be considered in the future as it would improve the tracking accuracy. More advanced tracking algorithms such as Particle Filters should also be considered.

The system currently tries to localise based on any signals it receives. Voice detection methods can be implemented to activate the localisation only when people are speaking rather than for random noises and sounds.

It will also be required to extend the localisation algorithm for multiple simultaneous speakers.

Appendix A: Derivation of Algorithms

A.1. Derivation of Spherical Intersection (SX) method

The SX method for localisation uses matrices to find the position of a source based on time-delays between source and receivers. This method for localisation uses the notion of a reference receiver so that time-delays can be compared to each other.

Let the sound source be at the unknown position (x, y) and the N sensors to be at the known coordinates (x_i, y_i) for i = 1 to N.

The squared distance between the source and receiver i is given by

$$\begin{aligned} r_i^2 &= (x_i - x)^2 + (y_i - y)^2 \\ &= x^2 + y^2 - 2x_i x - 2y_i y + x_i^2 + y_i^2 \end{aligned} \quad (31)$$

Let

$$K_i = x_i^2 + y_i^2$$

Then

$$r_i^2 = x^2 + y^2 - 2x_i x - 2y_i y + K_i \ . \quad (32)$$

Without any loss of generality, let sensor 1 be the reference sensor. For the case when i = 1

$$r_1^2 = x^2 + y^2 - 2x_1 x - 2y_1 y + K_1 \quad (33)$$

Let $d_{i,1}$ be the TDOA between sensor i and sensor 1 (the reference receiver) c be the speed of propagation of the signal and $r_{i,1}$ be the TDOA distance. Then

$$r_{i,1} = cd_{i,1} = r_i - r_1 \quad (34)$$

Rearranging (34) and squaring

$$\begin{aligned} r_i^2 &= (r_{i,1} + r_1)^2 \\ &= r_{i,1}^2 + 2r_{i,1}r_1 + r_1^2 \ . \end{aligned} \quad (35)$$

Equating (32) and (35)

$$r_{i,1}^2 + 2r_{i,1}r_1 + r_1^2 = x^2 + y^2 - 2x_i x - 2y_i y + K_i \quad (36)$$

Subtract (33) from (36)

$$r_{i,1}^2 + 2r_{i,1}r_1 = -2x_{i,1}x - 2y_{i,1}y + K_i - K_1 \tag{37}$$

where

$x_{i,1} = x_i - x_1, \qquad y_{i,1} = y_i - y_1.$

If this is done for each i then a matrix equation can be found for (x, y)

$$2X\begin{bmatrix} x \\ y \end{bmatrix} = 2Ar_1 + B \tag{38}$$

where

$$X = \begin{bmatrix} x_{2,1} & y_{2,1} \\ x_{3,1} & y_{3,1} \\ \vdots & \vdots \\ x_{N,1} & y_{N,1} \end{bmatrix}, \qquad A = -\begin{bmatrix} r_{2,1} \\ r_{3,1} \\ \vdots \\ r_{N,1} \end{bmatrix}, \quad B = -\begin{bmatrix} r_{2,1}^2 \\ r_{3,1}^2 \\ \vdots \\ r_{N,1}^2 \end{bmatrix} + \begin{bmatrix} K_2 - K_1 \\ K_3 - K_1 \\ \vdots \\ K_N - K_1 \end{bmatrix}.$$

Pre-multiplying both sides by X^T and dividing by 2, where X^T = transpose of X

$$X^T X\begin{bmatrix} x \\ y \end{bmatrix} = X^T\left(Ar_1 + \frac{1}{2}B\right) \tag{39}$$

Pre-multiplying both sides by the inverse of $X^T X$ gives a set of equations in the form of

$$\begin{bmatrix} x \\ y \end{bmatrix} = Cr_1 + D \tag{40}$$

where

$$C = \begin{bmatrix} C_x \\ C_2 \end{bmatrix} = \left(X^T X\right)^{-1} X^T A, \qquad D = \begin{bmatrix} D_x \\ D_y \end{bmatrix} = \left(X^T X\right)^{-1} X^T\left(\frac{1}{2}\right)B.$$

Note that the above equation can only be solved if the inverse of $X^T X$ exists. This requires X to have more than one row (which corresponds to having three or more microphones) and the rows of X to be linearly independent (which corresponds to the microphones not being collinear). Substituting (40) into (33) gives an equation in the form of

$$\alpha r_1^2 + \beta r_1 + \chi = 0 \tag{41}$$

where

$$\alpha = C_x^2 + C_y^2 - 1,$$
$$\beta = 2\left(C_x\left(D_x - x_1\right) + C_y\left(D_y - y_1\right)\right),$$
$$\chi = D_x\left(D_x - 2x_1\right) + D_y\left(D_y - 2y_1\right) + K_1.$$

Solving the quadratic (41) gives values for r_1 which can then be substituted into (40) to give an estimate for (x, y). The correct value to choose from the quadratic roots is discussed in Section 4.1.1.

A.2. Derivation of Ho-Chan Method

The Ho-Chan Methods starts with the same equations as the SX method and arrives at the same result as (38) which is repeated here:

$$2X\begin{bmatrix} x \\ y \end{bmatrix} = 2Ar_1 + B \tag{38}$$

where

$$X = \begin{bmatrix} x_{2,1} & y_{2,1} \\ x_{3,1} & y_{3,1} \\ \vdots & \vdots \\ x_{N,1} & y_{N,1} \end{bmatrix}, \qquad A = -\begin{bmatrix} r_{2,1} \\ r_{3,1} \\ \vdots \\ r_{N,1} \end{bmatrix}, \qquad B = -\begin{bmatrix} r_{2,1}^2 \\ r_{3,1}^2 \\ \vdots \\ r_{N,1}^2 \end{bmatrix} + \begin{bmatrix} K_2 - K_1 \\ K_3 - K_1 \\ \vdots \\ K_N - K_1 \end{bmatrix}.$$

Arranging this equation:

$$X\begin{bmatrix} x \\ y \end{bmatrix} - Ar_1 = \frac{1}{2}B \tag{42}$$

.

Merging the left side of (42) and re-arranging becomes

$$-\frac{1}{2}B + \begin{bmatrix} X & -A \end{bmatrix}\begin{bmatrix} x \\ y \\ r_1 \end{bmatrix} = 0 \tag{43}$$

.

Based on (43) define

$$\mathbf{h} - \mathbf{G}_a\mathbf{z}_a^0 = 0 \tag{44}$$

where

$$\mathbf{h} = -\frac{1}{2}B = \frac{1}{2}\begin{bmatrix} r_{2,1}^2 - K_2 + K_1 \\ r_{3,1}^2 - K_3 + K_1 \\ \vdots \\ r_{N,1}^2 - K_N + K_1 \end{bmatrix}, \qquad \mathbf{G}_a = -\begin{bmatrix} X & -A \end{bmatrix} = -\begin{bmatrix} x_{2,1} & y_{2,1} & r_{2,1} \\ x_{3,1} & y_{3,1} & r_{3,1} \\ \vdots & \vdots & \vdots \\ x_{N,1} & y_{N,1} & r_{N,1} \end{bmatrix},$$

$$\mathbf{z}_a{}^0 = \begin{bmatrix} x^0 \\ y^0 \\ r_1{}^0 \end{bmatrix}.$$

x^0 denotes the true value of x. (44) only holds when $\mathbf{h}$ and $\mathbf{G}_a$ contain the true TDOA distances $r_{i,1}$ for i = 2 to N. In practice the true TDOA distances are not available. The error vector ψ is thus defined as

$$\mathbf{h} - \mathbf{G}_a\mathbf{z}_a{}^0 = \psi \tag{45}$$

where

$$\psi = \begin{bmatrix} \psi_2 \\ \psi_3 \\ \vdots \\ \psi_N \end{bmatrix}.$$

The TDOA distances are given as

$$r_{i,1} = r_{i,1}^0 + n_{i,1}, \qquad i = 2 \text{ to } N \tag{46}$$

where $r_{i,1}^0$ are the true TDOA distances and $n_{i,1}$ are the errors in the TDOA distances.

From the definition of TDOA distances

$$r_{i,1}^0 = r_i^0 - r_1^0. \tag{47}$$

Substituting (46) and (47) into (45)

$$\begin{aligned} \psi_i &= \frac{1}{2}\left(r_{i,1}{}^2 - K_i + K_1\right) + x_{i,1}x^0 + y_{i,1}y^0 + r_{i,1}r_1^0 \\ &= \frac{1}{2}\left(r_{i,1}^{0\,2} + 2r_{i,1}^0 n_{i,1} + n_{i,1}{}^2 - K_i + K_1\right) + x_{i,1}x^0 + y_{i,1}y^0 + r_{i,1}^0 r_1^0 + r_1^0 n_{i,1}. \end{aligned}$$

Recall from (44) that with true TDOA distances

$$0 = \frac{1}{2}\left(r_{i,1}^{0\,2} - K_i + K_1\right) + x_{i,1}x^0 + y_{i,1}y^0 + r_{i,1}^0 r_1^0$$

hence

$$\begin{aligned}\psi_i &= \frac{1}{2}\left(2r_{i,1}^0 n_{i,1} + n_{i,1}^{\,2}\right) + r_1^0 n_{i,1} \\ &= \frac{1}{2}\left(2r_i^0 n_{i,1} - 2r_1^0 n_{i,1} + n_{i,1}^{\,2}\right) + r_1^0 n_{i,1} \\ &= r_i^0 n_{i,1} + \frac{1}{2} n_{i,1}^{\,2}. \end{aligned} \tag{48}$$

Therefore

$$\psi = \mathbf{B}\mathbf{n} + \frac{1}{2}\mathbf{n}\odot\mathbf{n} \tag{49}$$

where

$$\mathbf{B} = \begin{bmatrix} r_2^0 & 0 & \cdots & 0 \\ 0 & r_3^0 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & r_N^0 \end{bmatrix}, \qquad \mathbf{n} = \begin{bmatrix} n_{2,1} \\ n_{3,1} \\ \vdots \\ n_{N,1} \end{bmatrix}.$$

n is the error vector of the TDOA distances; the symbol ⊙ represents the Schur product (element-by-element product).

The covariance matrix **Ψ** of ψ is evaluated with the condition that the signal-to-noise ratio (SNR) is high, i.e. $n_{i,1} << r_i^0$. The second term on the right of (49) is ignored and the covariance matrix is given by

$$\mathbf{\Psi} = \mathbf{E}[\psi\,\psi^T] = \mathbf{B}\mathbf{Q}\mathbf{B} \tag{50}$$

where $\mathbf{E}[x]$ is the expected value of x and **Q** is the covariance matrix of **n**.

The set of equations in (45) are solved by LS with the assumption that the elements in $\mathbf{z}_a$ have no relationship:

$$\mathbf{z}_a = (\mathbf{G}_a^T\mathbf{\Psi}^{-1}\mathbf{G}_a)^{-1}\mathbf{G}_a^T\mathbf{\Psi}^{-1}\mathbf{h} \tag{51}$$

In practice $\mathbf{\Psi}$ is not known since $\mathbf{B}$ contains the true TDOA distances. The Ho-Chan method uses further approximation to estimate $\mathbf{B}$. When the source is far from the sensor array, each r_i^0 is close to r^0 so that $\mathbf{B} \approx r^0\mathbf{I}$, where r^0 is the range of the source and $\mathbf{I}$ is the identity matrix of size N-1. Scaling of $\mathbf{\Psi}$ does not affect the answer, so an approximation of (51) is

$$\mathbf{z}_a \approx (\mathbf{G}_a^T\mathbf{Q}^{-1}\mathbf{G}_a)^{-1}\mathbf{G}_a^T\mathbf{Q}^{-1}\mathbf{h} \tag{52}$$

For the case where the source is close to the sensor array (52) is used to obtain an initial estimate of $\mathbf{z}_a$ which is used to find an estimate of $\mathbf{B}$. This is then put into (51). (51) can be iterated multiple times to produce an even better estimate, but simulations done by Chan and Ho (1994) show that one iteration is sufficient to produce accurate results.

The covariance of $\mathbf{z}_a$ is given by

$$\text{cov}(\mathbf{z}_a) = E[\Delta\mathbf{z}_a\Delta\mathbf{z}_a^T] = (\mathbf{G}_a^{0T}\mathbf{\Psi}^{-1}\mathbf{G}_a^0)^{-1} \tag{53}$$

Where $\mathbf{G}_a^0$ is identical to $\mathbf{G}_a$, as defined in (44), except that the TDOA values, $r_{2,1}$ $r_{3,1}$ .. $r_{N,1}$ are exact, ie.

$$G_a^0 = \begin{bmatrix} x_{2,1} & y_{2,1} & r_{2,1}^0 \\ x_{3,1} & y_{3,1} & r_{3,1}^0 \\ \vdots & \vdots & \vdots \\ x_{N,1} & y_{N,1} & r_{N,1}^0 \end{bmatrix}$$

.

The above solution of $\mathbf{z}_a$ assumes that x, y and r_1 are independent. However they are related by

$$r_1^2 = (x_1 - x)^2 + (y_1 - y)^2 \tag{54}$$

Let the elements of $\mathbf{z}_a$ be expressed as

$$\mathbf{z}_a = \begin{bmatrix} z_{a,1} \\ z_{a,2} \\ z_{a,3} \end{bmatrix} = \begin{bmatrix} x^0 + e_1 \\ y^0 + e_2 \\ r_1^0 + e_3 \end{bmatrix} \tag{55}$$

where e_1, e_2 and e_3 are estimation errors of $\mathbf{z}_a$.

Define another set of equations:

$$\mathbf{h}' - \mathbf{G}_a'\mathbf{z}_a'^0 = \psi' \tag{56}$$

where

$$\mathbf{h}' = \begin{bmatrix} (z_{a,1} + x_1)^2 \\ (z_{a,2} + y_1)^2 \\ z_{a,3}{}^2 \end{bmatrix}, \mathbf{G}_a' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{z}_a' = \begin{bmatrix} (x - x_1)^2 \\ (y - y_1)^2 \end{bmatrix}.$$

$\psi' = \begin{bmatrix} \psi_1' \\ \psi_2' \\ \psi_3' \end{bmatrix}$ is the vector of inaccuracies in $\mathbf{z}_a$. Substituting (56) into (55) gives

$$\begin{aligned} \psi_1' &= 2(x^0 - x_1)e_1 + e_1^2 \\ \psi_2' &= 2(y^0 - y_1)e_2 + e_2^2 \\ \psi_3' &= 2r_1^0 e_3 + e_3^2. \end{aligned} \tag{57}$$

For small errors (57) can be approximated by

$$\begin{aligned} \psi_1' &\approx 2(x^0 - x_1)e_1 \\ \psi_2' &\approx 2(y^0 - y_1)e_2 \\ \psi_3' &\approx 2r_1^0 e_3. \end{aligned} \tag{58}$$

The covariance matrix $\mathbf{\Psi}'$ of ψ' is

$$\mathbf{\Psi}' = \mathbf{E}[\psi'\psi'^T] = 4\mathbf{B}'\mathrm{cov}(\mathbf{z}_a)\mathbf{B}' \tag{59}$$

where

$$\mathbf{B}' = \begin{bmatrix} (x^0 - x_1) & 0 & 0 \\ 0 & (y^0 - y_1) & 0 \\ 0 & 0 & r_1^0 \end{bmatrix}$$

$\mathbf{B}'$ can be approximated by using the values in $\mathbf{z}_a$. A second LS calculation is used to find $\mathbf{z}_a'$

$$\mathbf{z}_a' = (\mathbf{G}_a'^T\mathbf{\Psi}'^{-1}\mathbf{G}_a')^{-1}\mathbf{G}_a'^T\mathbf{\Psi}'^{-1}\mathbf{h}' \tag{60}$$

The Covariance matrix of $\mathbf{z}_a'$ is

$$\mathrm{cov}(\mathbf{z}_a') = (\mathbf{G}_a'^T\mathbf{\Psi}'^{-1}\mathbf{G}_a')^{-1} \tag{61}$$

Since $\mathbf{z}_a'$ is an estimate of $(x - x_1)^2$ and $(y - y_1)^2$ a simple conversion will produce x and y estimates:

$$\mathbf{z}_p = \pm\sqrt{\mathbf{z}_a'} + \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \tag{62}$$

Since there are two possible solutions for a square root, the correct solution is the one that lies in the region of interest. This can be easily determined by looking at the sign of the initial $\mathbf{z}_a$ estimates. If one of the coordinates is close to zero the square root may become imaginary. In this case the imaginary component should be set to zero (Chan and Ho 1994).

The covariance matrix $\mathbf{\Phi}$ of $\mathbf{z}_p$ is

$$\mathbf{\Phi} = \text{cov}(\mathbf{z}_p) = \frac{1}{4}\mathbf{B}''^{-1}\text{cov}(\mathbf{z}_a')\mathbf{B}''^{-1} \tag{63}$$

where

$$\mathbf{B}'' = \begin{bmatrix} (x^0 - x_1) & 0 \\ 0 & (y^0 - y_1) \end{bmatrix}.$$

To summarise, (52) is used to find an initial estimate of $\mathbf{z}_a$ which is used to estimate $\mathbf{B}$. (51) is then used to find more accurate estimates of $\mathbf{z}_a$ and can be repeated for multiple iterations. (60) is then used to put the relationship between x, y and r_1 into the calculation. The final coordinate estimate $\mathbf{z}_p$ is found using (62).

A.3. Derivation of Cramér-Rao Lower Bound (CRLB) for Localisation

The CRLB for localisation is based on Gromov, Akos, Pullen, Enge and Parkinson (2000). Gromov et al. (2000) give a CRLB calculation for 3D localisation that uses radius distance, elevation angle and azimuth angle errors of estimated source coordinates from the reference (0, 0, 0) point. The CRLB Error Covariance Matrix (ECM) is defined by

$$B_{loc} = \left(H^T B_{\Delta R_i}^{-1} H\right)^{-1} \tag{64}$$

where

$$H = \begin{bmatrix} \frac{\partial \Delta R_1}{\partial R} & \frac{\partial \Delta R_1}{\partial \beta} & \frac{\partial \Delta R_1}{\partial \varepsilon} \\ \frac{\partial \Delta R_2}{\partial R} & \frac{\partial \Delta R_2}{\partial \beta} & \frac{\partial \Delta R_2}{\partial \varepsilon} \\ \vdots & \vdots & \vdots \\ \frac{\partial \Delta R_{N-1}}{\partial R} & \frac{\partial \Delta R_{N-1}}{\partial \beta} & \frac{\partial \Delta R_{N-1}}{\partial \varepsilon} \end{bmatrix}, \; R = \text{radius}, \; \beta = \text{azimuth}, \; \varepsilon = \text{elevation},$$

$B_{\Delta R_i}$ = ECM of the TDOA distances.

ΔR_i are the TDOA distances for i = 1 to N-1, N is the number of sensors. This was derived using cosine and sine rules of triangles:

$$\Delta R_i = R\left(1-\sqrt{1-\frac{2l_i}{R}\left(\cos\varepsilon\cos\varepsilon_i\cos(\beta-\beta_i)+\sin\varepsilon\sin\varepsilon_i\right)+\frac{l_i^2}{R^2}}\right) \qquad (65)$$

where l_i, β_i and ε_i are the radius distance, azimuth and elevation angles of sensor i for i = 1 to N.

Using the same principle the CRLB of 2D Cartesian coordinate localisation was derived. Taking the square root of (31) and (54):

$$r_i = \sqrt{(x_i - x)^2 + (y_i - y)^2}$$

$$r_1 = \sqrt{(x_1 - x)^2 + (y_1 - y)^2}$$

the TDOA distances are given by

$$r_{i,1} = r_i - r_1 = \sqrt{(x_i - x)^2 + (y_i - y)^2} - \sqrt{(x_1 - x)^2 + (y_1 - y)^2}\ . \qquad (66)$$

Taking the partial derivatives

$$\frac{\partial r_{i,1}}{\partial x} = \frac{x_i - x}{\sqrt{(x_i - x)^2 + (y_i - y)^2}} - \frac{x_1 - x}{\sqrt{(x_1 - x)^2 + (y_1 - y)^2}}$$

$$\frac{\partial r_{i,1}}{\partial y} = \frac{y_i - y}{\sqrt{(x_i - x)^2 + (y_i - y)^2}} - \frac{y_1 - y}{\sqrt{(x_1 - x)^2 + (y_1 - y)^2}}$$

the CRLB ECM is thus

$$ECM = \left(H^T B_{r_{i,1}}^{-1} H\right)^{-1} \qquad (67)$$

where H is

$$H = \begin{bmatrix} \dfrac{\partial r_{2,1}}{\partial x} & \dfrac{\partial r_{2,1}}{\partial y} \\ \dfrac{\partial r_{3,1}}{\partial x} & \dfrac{\partial r_{3,1}}{\partial y} \\ \vdots & \vdots \\ \dfrac{\partial r_{N,1}}{\partial x} & \dfrac{\partial r_{N,1}}{\partial y} \end{bmatrix}.$$

The ECM is a 2x2 matrix with the diagonals containing lower bound variances for x and y coordinates in the first and second diagonal elements respectively. Summing of these will result in the lower bound variance of the distance used in the CRLB plots in Section 4.1 and Appendix B.

The CRLB for 3D localisation is a straight extension of the 2D case:

$$r_i = \sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2}$$

$$r_1 = \sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2}$$

$$r_{i,1} = \sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2} - \sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2}, \qquad (68)$$

$$\frac{\partial r_{i,1}}{\partial x} = \frac{x_i - x}{\sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2}} - \frac{x_1 - x}{\sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2}}$$

$$\frac{\partial r_{i,1}}{\partial y} = \frac{y_i - y}{\sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2}} - \frac{y_1 - y}{\sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2}}$$

$$\frac{\partial r_{i,1}}{\partial z} = \frac{z_i - z}{\sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2}} - \frac{z_1 - z}{\sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2}}. \qquad (69)$$

The CRLB is

$$ECM = \left(H^T B_{r_{i,1}}^{-1} H\right)^{-1} \qquad (70)$$

where

$$H = \begin{bmatrix} \frac{\partial r_{2,1}}{\partial x} & \frac{\partial r_{2,1}}{\partial y} & \frac{\partial r_{2,1}}{\partial z} \\ \frac{\partial r_{2,1}}{\partial x} & \frac{\partial r_{2,1}}{\partial y} & \frac{\partial r_{2,1}}{\partial z} \\ \vdots & \vdots & \vdots \\ \frac{\partial r_{N,1}}{\partial x} & \frac{\partial r_{N,1}}{\partial y} & \frac{\partial r_{N,1}}{\partial z} \end{bmatrix}.$$

The resulting ECM is a 3x3 matrix with the diagonals containing the lower bound variance of x, y and z coordinates. As for the 2D CRLB, summing of these values results in the distance lower bound variance.

Appendix B: Standard Deviation Plots

These standard deviation plots were created using TDOA distances with additive noise of set variances. There are four sets of plots which correspond to the two room sizes and the two noise variances used. Note that the colour bar scale on the side of each figure changes scale depending on the range of the values. Table 12 shows the different sets of plots.

Table 12. Sets of standard deviation plots.

Set number	Room size	Noise variance
1	4x4 m	0.01 m^2
2	4x4 m	0.0001 m^2
3	8x4 m	0.01 m^2
4	8x4 m	0.0001 m^2

Figures 21-30 show standard deviation plots for Set 1.

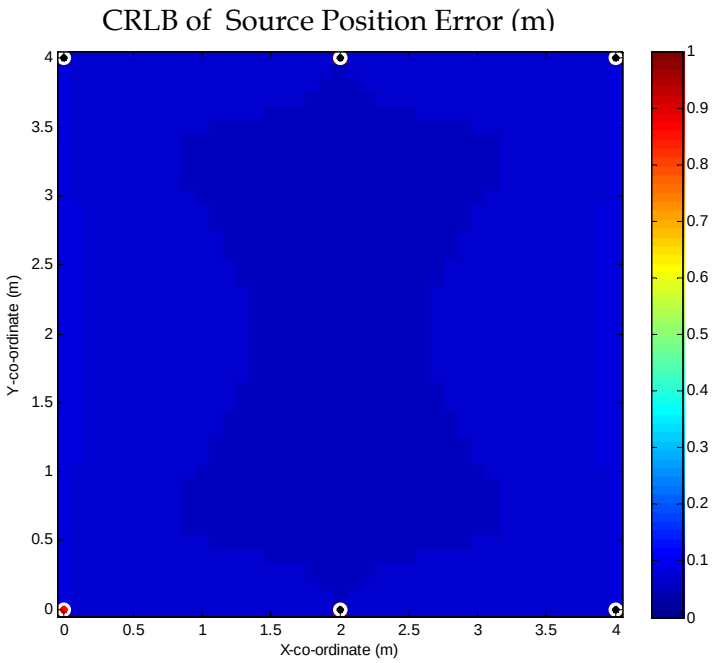

Figure 21. Standard deviation plot of CRLB of TDOA distance error; room size = 4x4 m; noise variance = 0.01 m2; reference sensor on the side.

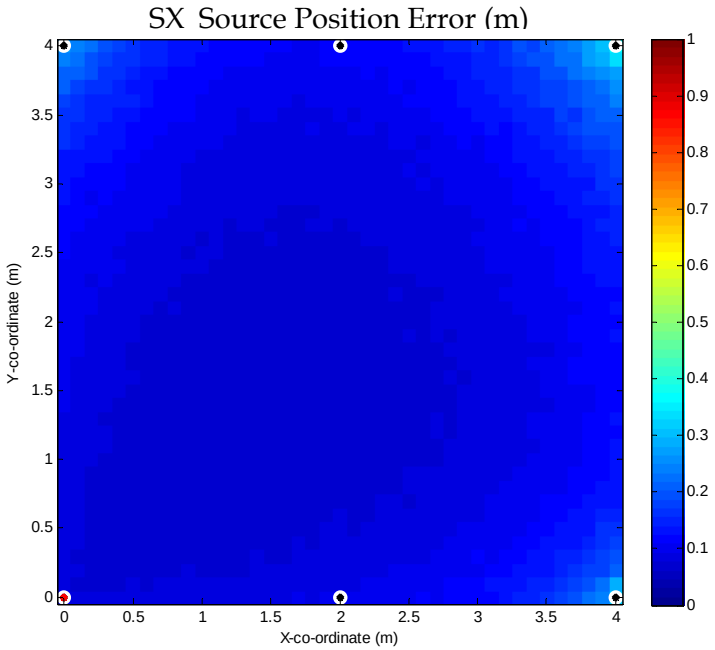

Figure 22. Standard deviation plot of SX method over 500 samples; room size = 4x4 m; noise variance = 0.01 m²; reference sensor on the side.

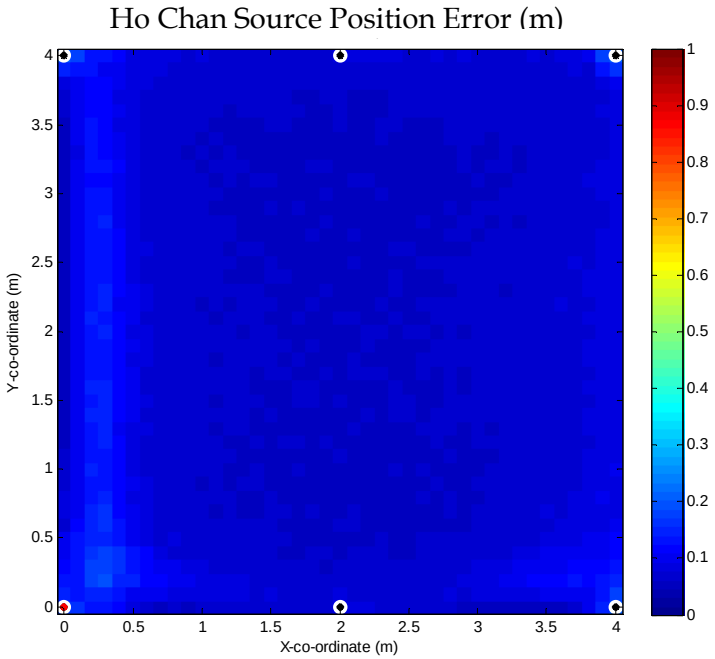

Figure 23. Standard deviation plot of Ho-Chan method over 500 samples; room size = 4x4 m; noise variance = 0.01 m²; reference sensor on the side.

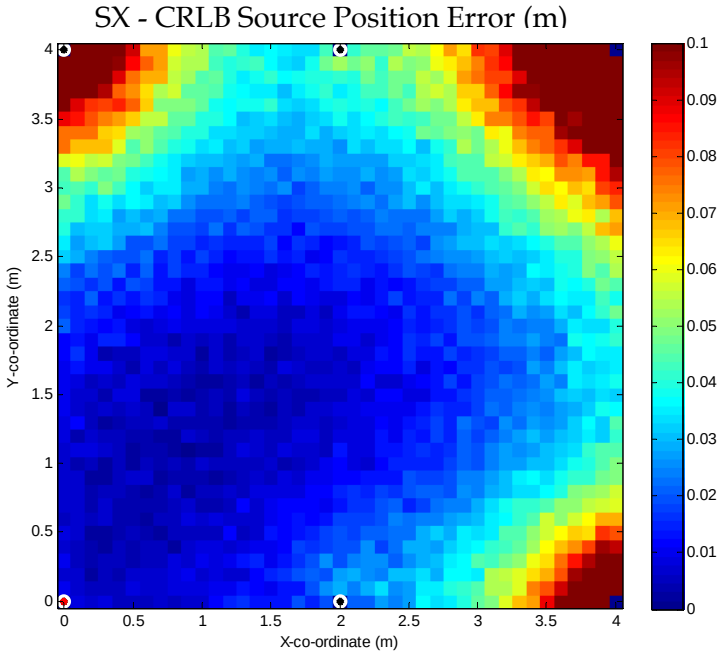

Figure 24. Plot of CRLB subtracted from SX standard deviations; room size = 4x4 m; noise variance = 0.01 m²; reference sensor on the side.

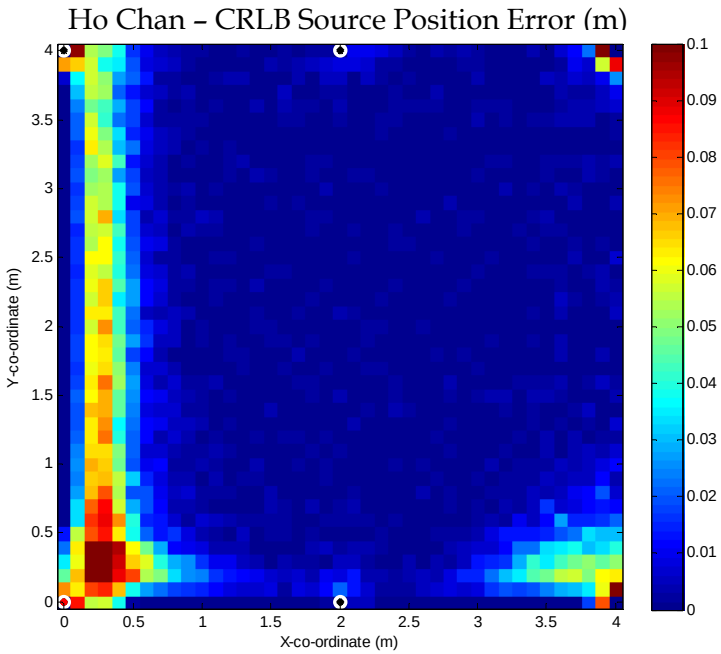

Figure 25. Plot of CRLB subtracted from Ho-Chan standard deviations; room size = 4x4 m; noise variance = 0.01 m²; reference sensor on the side.

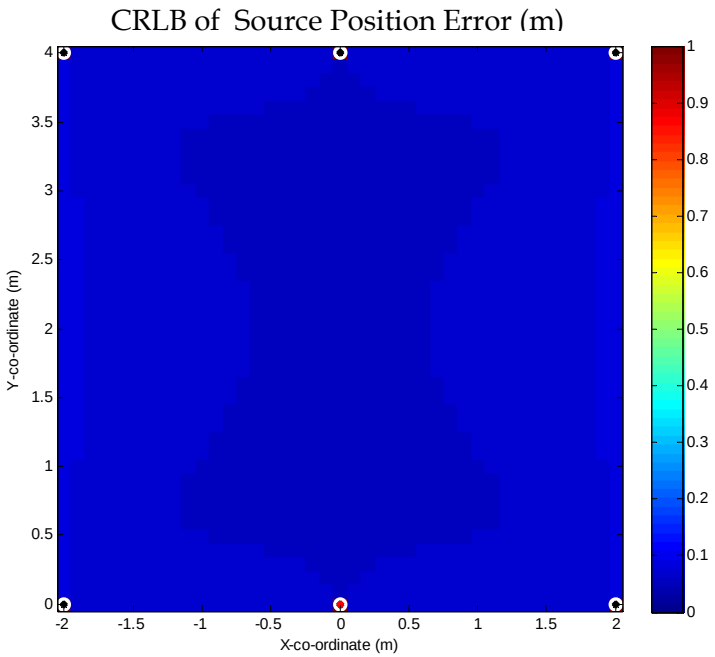

Figure 26. Standard deviation plot of CRLB of TDOA distance error; room size = 4x4 m; noise variance = 0.01 m²; reference sensor in the middle.

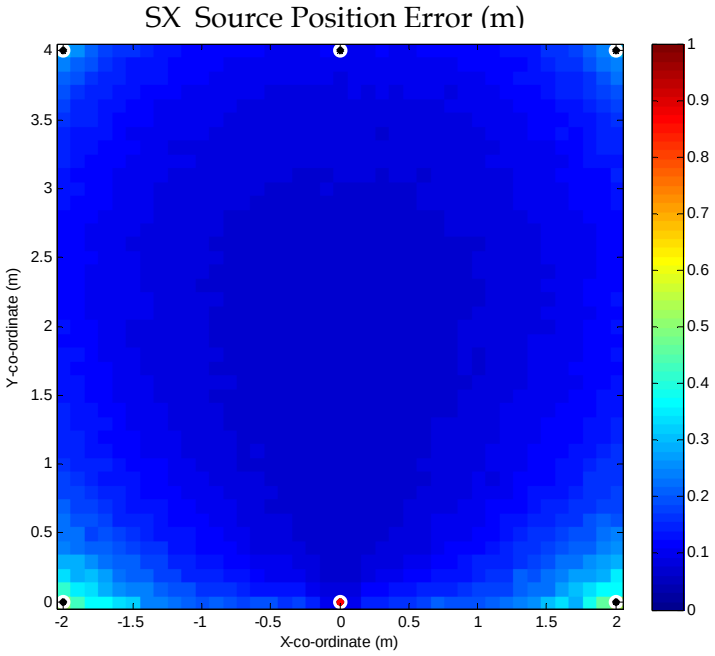

Figure 27. Standard deviation plot of SX method over 500 samples; room size = 4x4 m; noise variance = 0.01 m²; reference sensor in the middle.

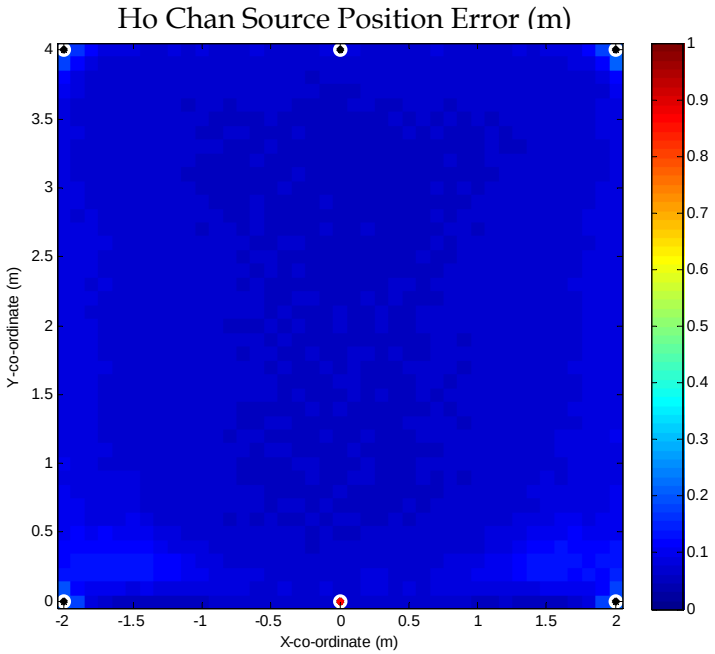

Figure 28. Standard deviation plot of Ho-Chan method over 500 samples; room size = 4x4 m; noise variance = 0.01 m²; reference sensor in the middle.

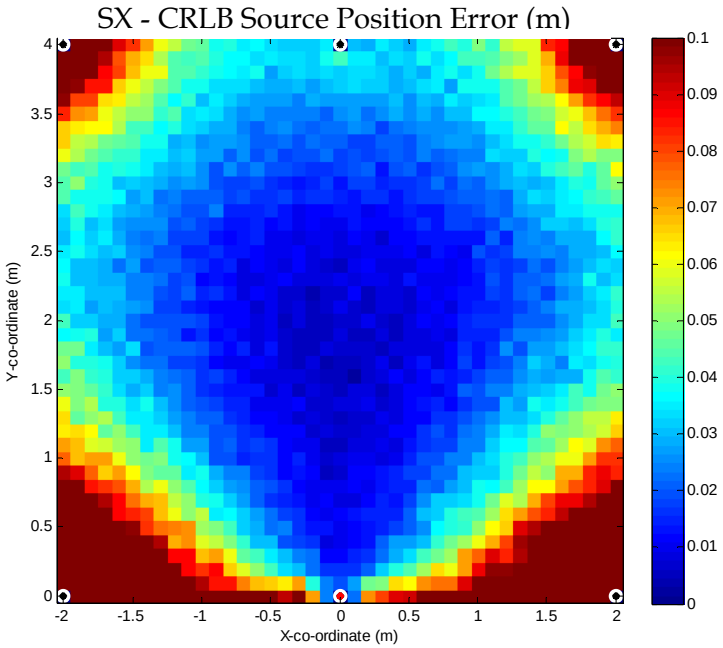

Figure 29. Plot of CRLB subtracted from SX standard deviations; room size = 4x4 m; noise variance = 0.01 m²; reference sensor in the middle.

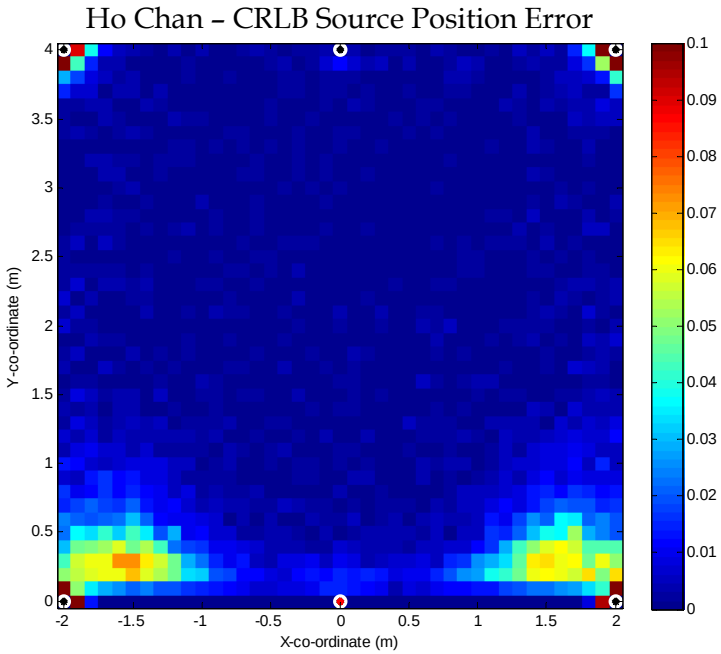

Figure 30. Plot of CRLB subtracted from Ho-Chan standard deviations; room size = 4x4 m; noise variance = 0.01 m²; reference sensor in the middle.

Figures 31-40 show the standard deviation plots of Set 2.

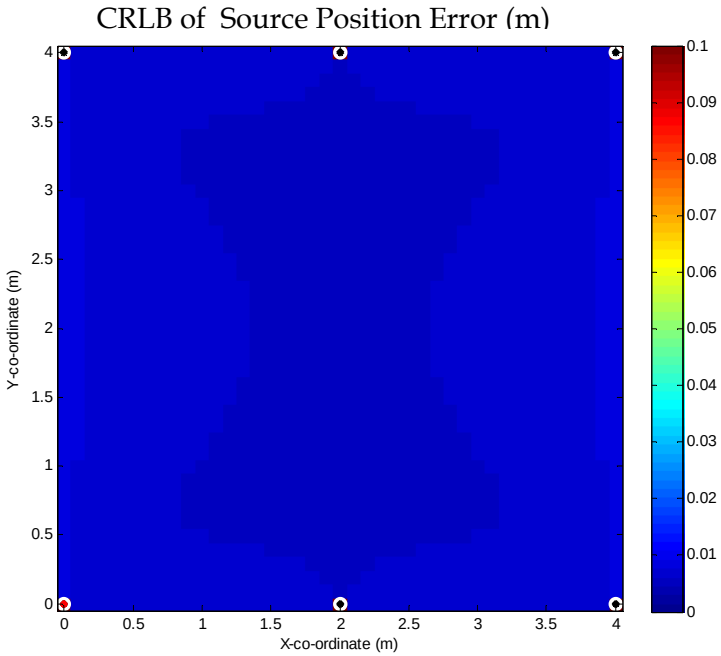

Figure 31. Standard deviation plot of CRLB of TDOA distance error; room size = 4x4 m; noise variance = 0.0001 m²; reference sensor on the side.

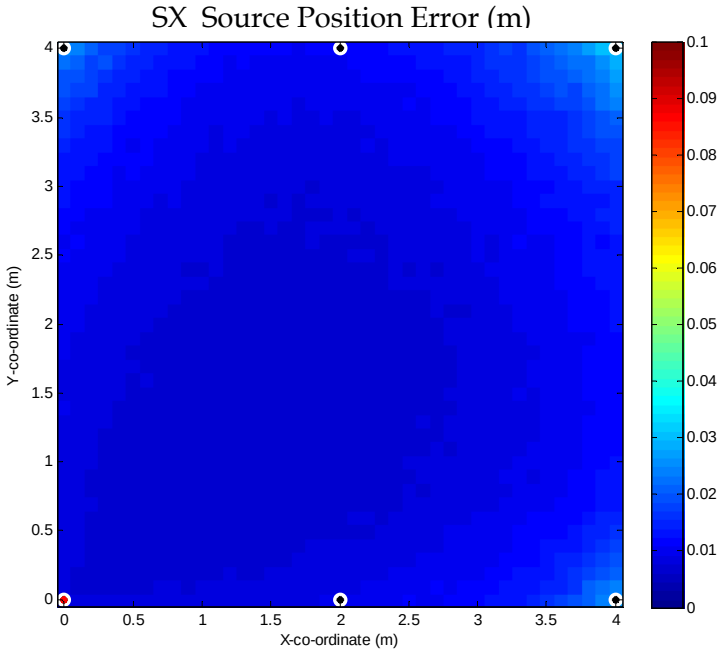

Figure 32. Standard deviation plot of SX method over 500 samples; room size = 4x4 m; noise variance = 0.0001 m²; reference sensor on the side.

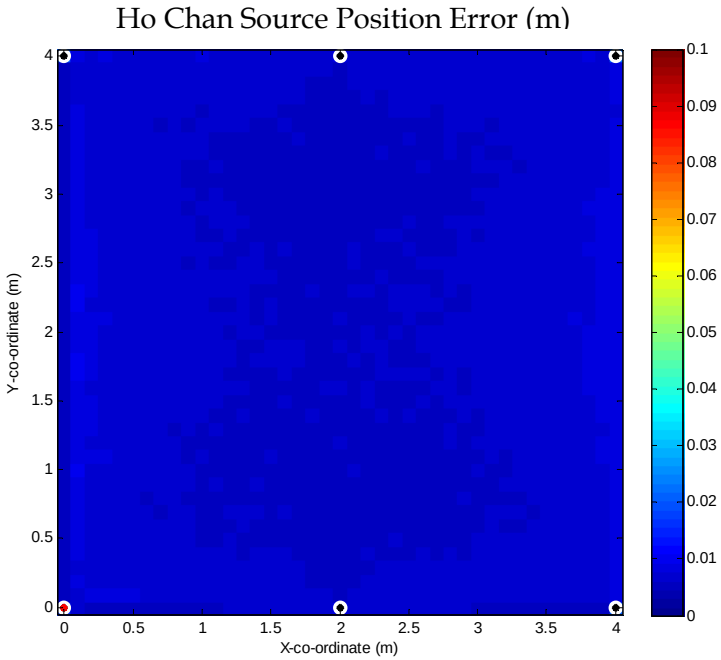

Figure 33. Standard deviation plot of Ho-Chan method over 500 samples; room size = 4x4 m; noise variance = 0.0001 m²; reference sensor on the side.

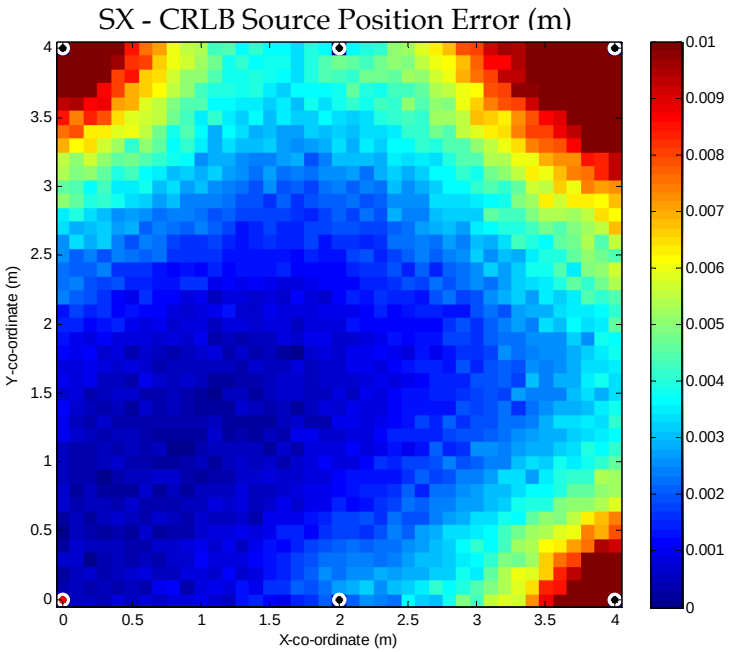

Figure 34. Plot of CRLB subtracted from SX standard deviations; room size = 4x4 m; noise variance = 0.0001 m²; reference sensor on the side.

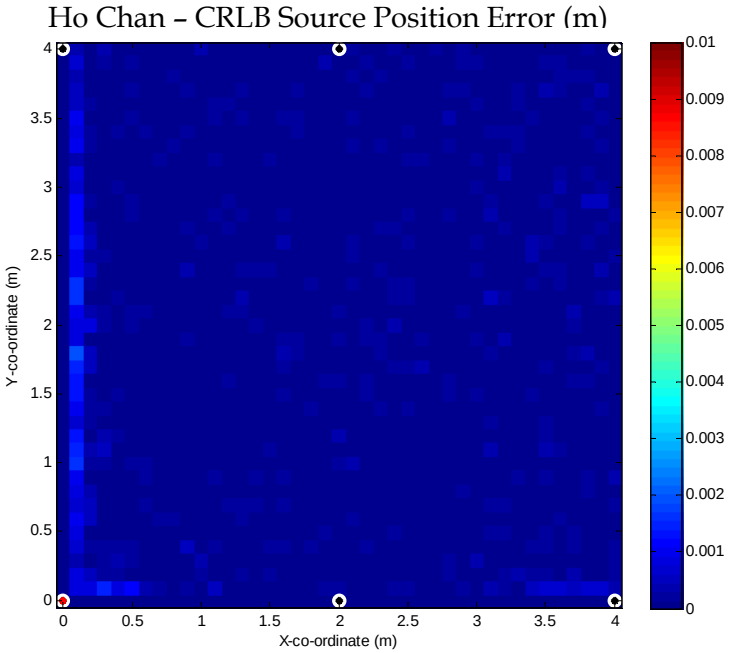

Figure 35. Plot of CRLB subtracted from Ho-Chan standard deviations; room size = 4x4 m; noise variance = 0.0001 m²; reference sensor on the side.

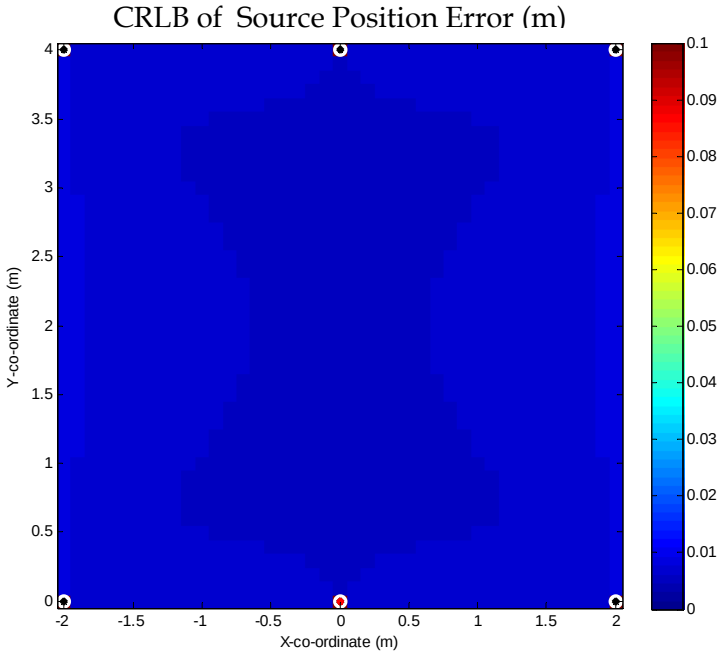

Figure 36. Standard deviation plot of CRLB of TDOA distance error; room size = 4x4 m; noise variance = 0.0001 m²; reference sensor in the middle.

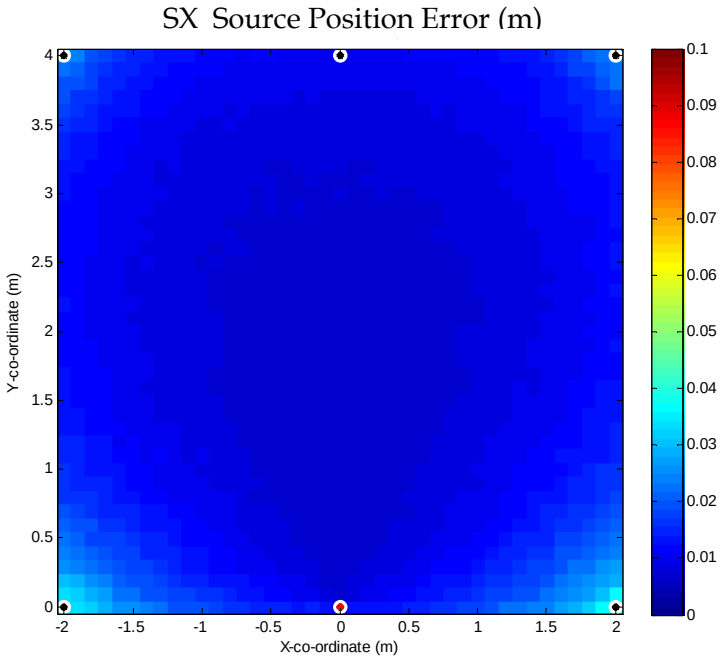

Figure 37. Standard deviation plot of SX method over 500 samples; room size = 4x4 m; noise variance = 0.0001 m²; reference sensor in the middle.

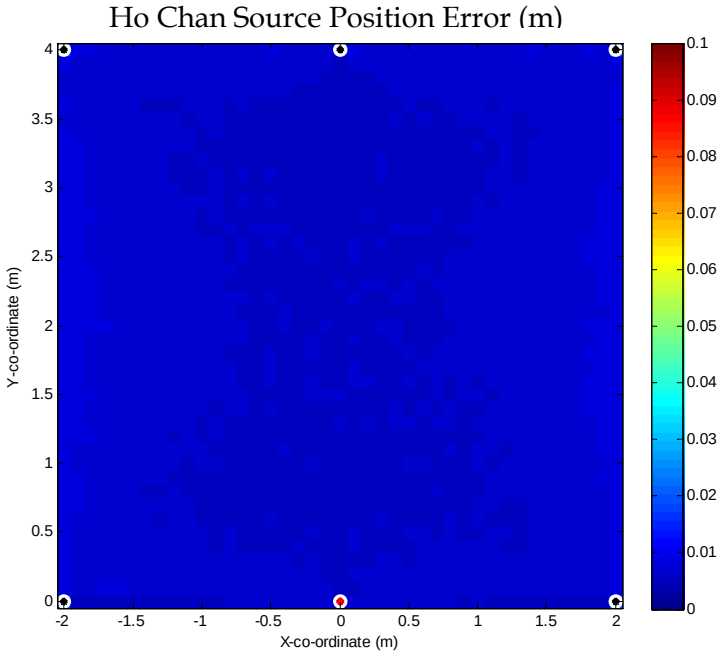

Figure 38. Standard deviation plot of Ho-Chan method over 500 samples; room size = 4x4 m; noise variance = 0.0001 m²; reference sensor in the middle.

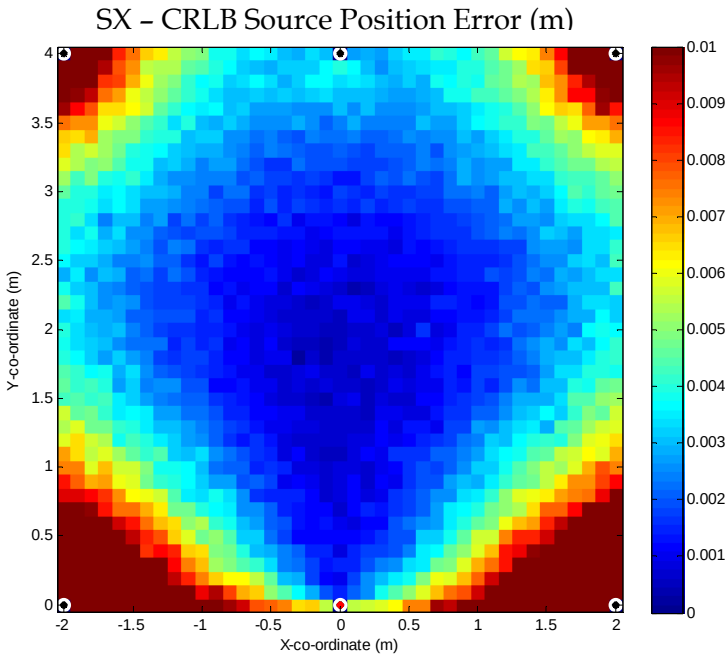

Figure 39. Plot of CRLB subtracted from SX standard deviations; room size = 4x4 m; noise variance = 0.0001 m²; reference sensor in the middle.

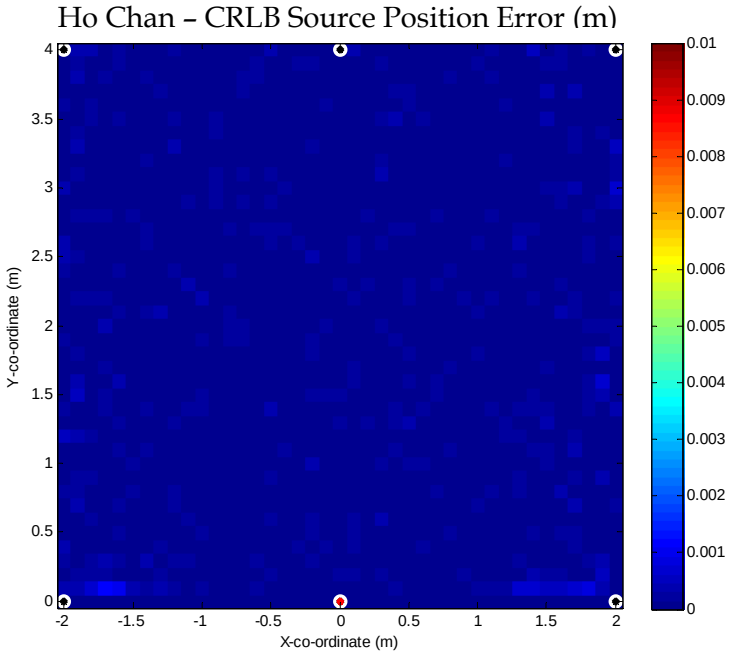

Figure 40. Plot of CRLB subtracted from Ho-Chan standard deviations; room size = 4x4 m; noise variance = 0.0001 m²; reference sensor in the middle.

Figures 41-50 show the standard deviation plots of Set 3.

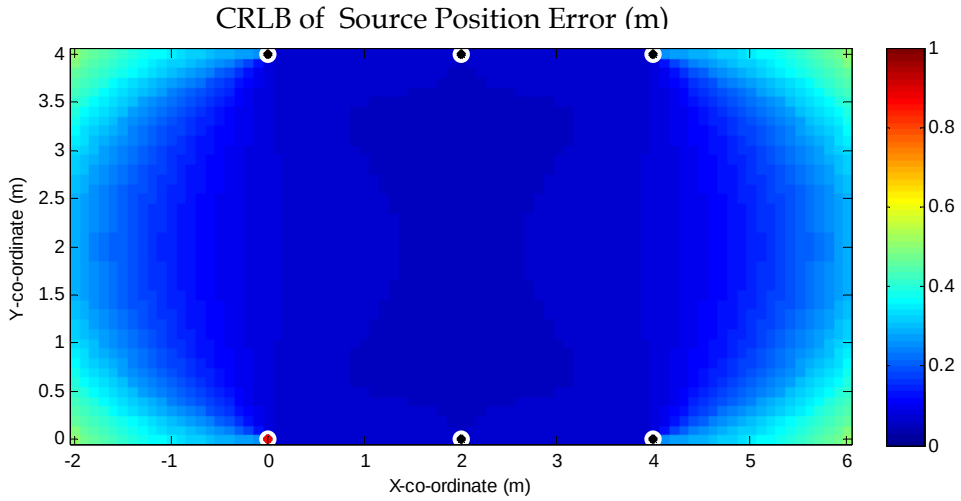

Figure 41. Standard deviation plot of CRLB of TDOA distance error; room size = 8x4 m; noise variance = 0.01 m²; reference sensor on the side.

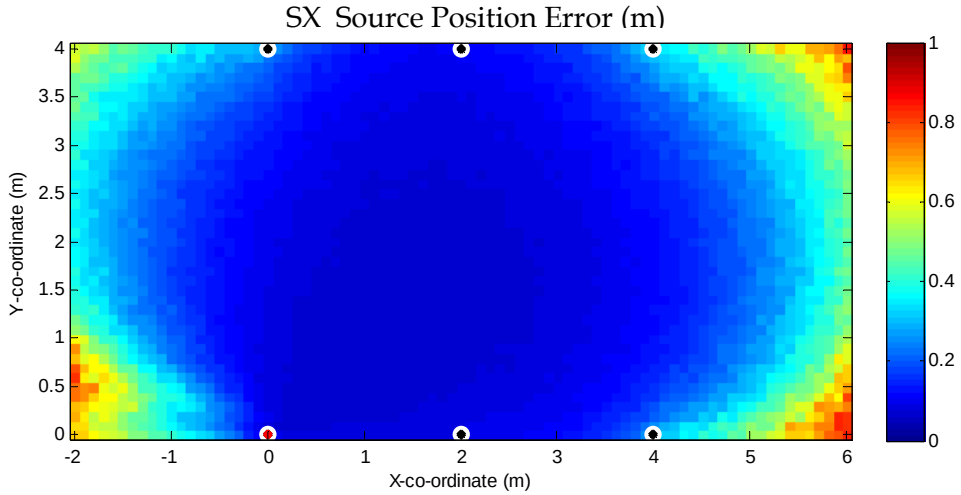

Figure 42. Standard deviation plot of SX method over 500 samples; room size = 8x4 m; noise variance = 0.01 m²; reference sensor on the side.

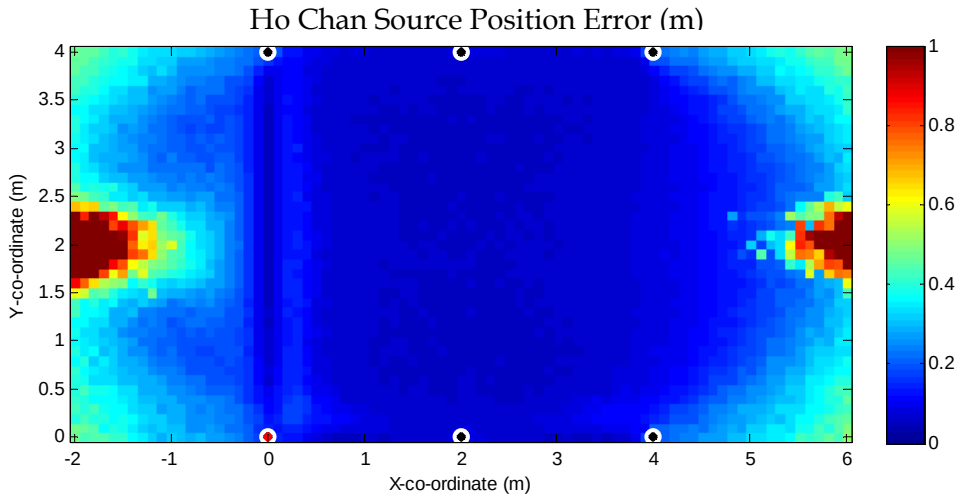

Figure 43. Standard deviation plot of Ho-Chan method over 500 samples; room size = 8x4 m; noise variance = 0.01 m²; reference sensor on the side.

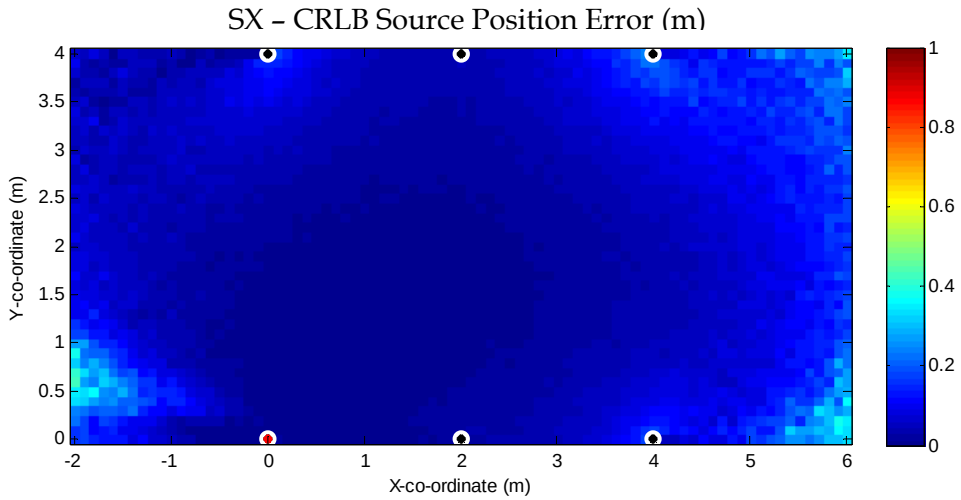

Figure 44. Plot of CRLB subtracted from SX standard deviations; room size = 8x4 m; noise variance = 0.01 m²; reference sensor on the side.

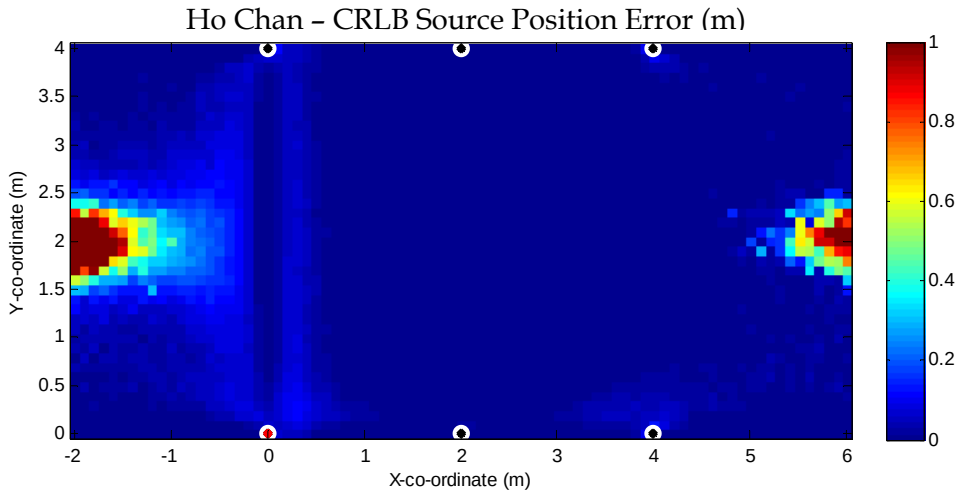

Figure 45. Plot of CRLB subtracted from Ho-Chan standard deviations; room size = 8x4 m; noise variance = 0.01 m²; reference sensor on the side.

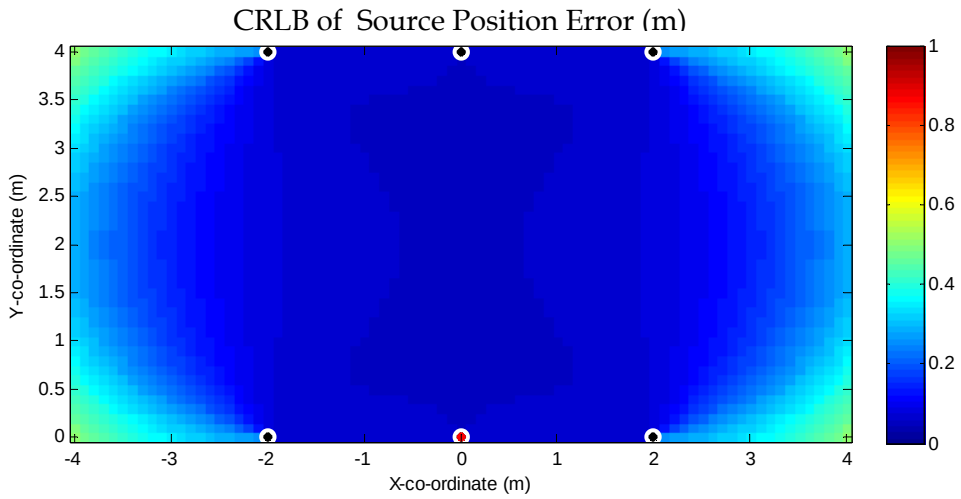

Figure 46. Standard deviation plot of CRLB of TDOA distance error; room size = 8x4 m; noise variance = 0.01 m²; reference sensor in the middle.

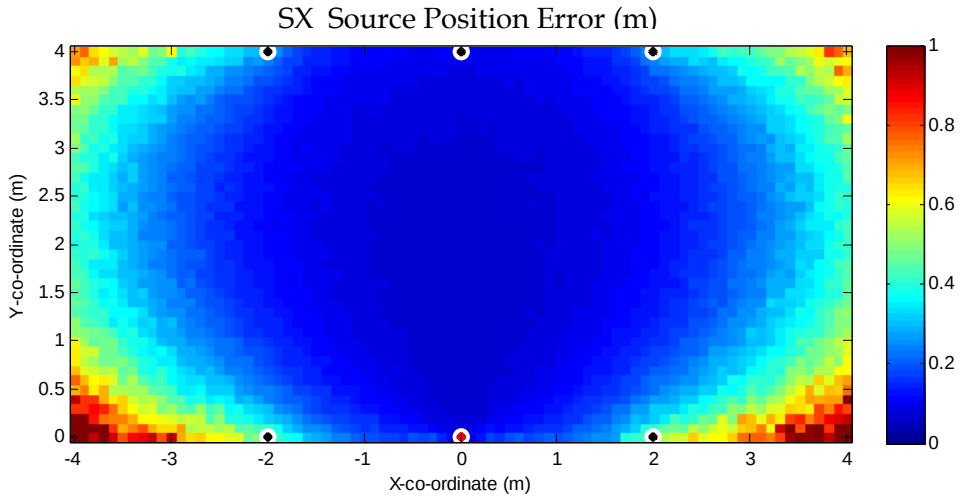

Figure 47. Standard deviation plot of SX method over 500 samples; room size = 8x4 m; noise variance = 0.01 m²; reference sensor in the middle.

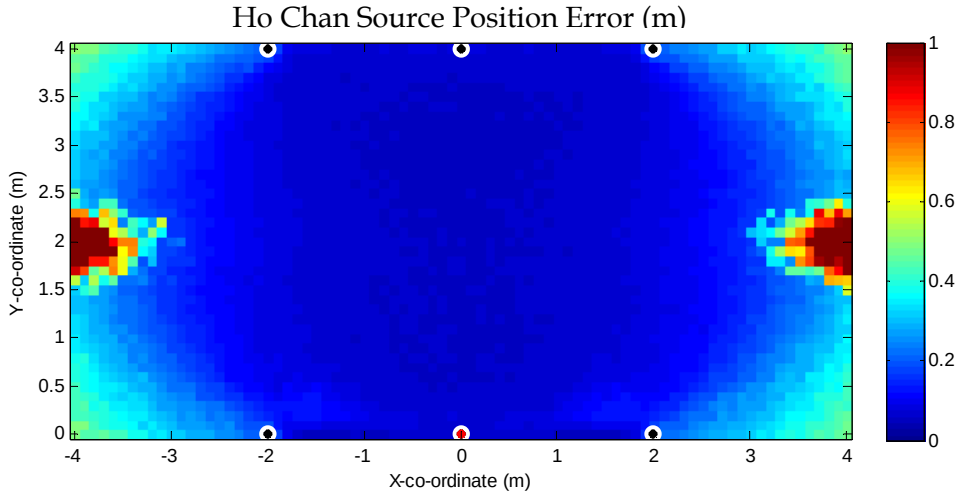

Figure 48. Standard deviation plot of Ho-Chan method over 500 samples; room size = 8x4 m; noise variance = 0.01 m²; reference sensor in the middle.

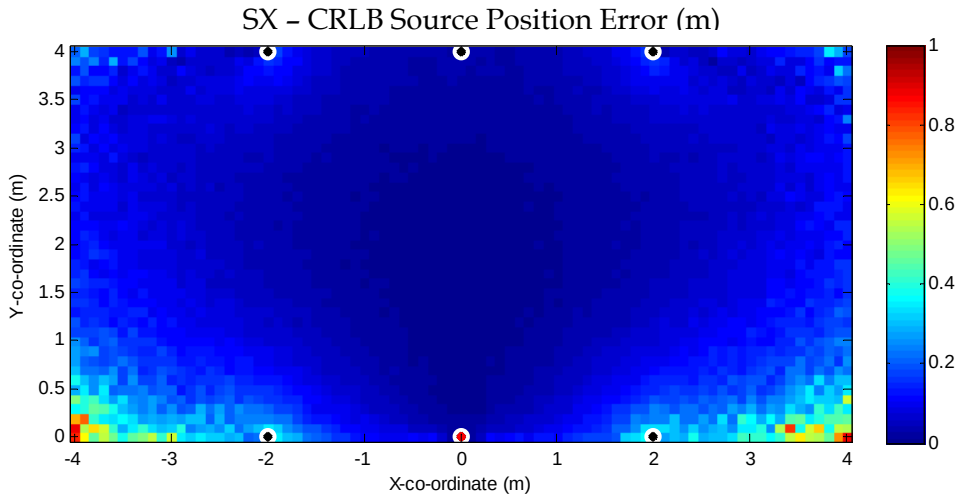

Figure 49. Plot of CRLB subtracted from SX standard deviations; room size = 8x4 m; noise variance = 0.01 m²; reference sensor in the middle.

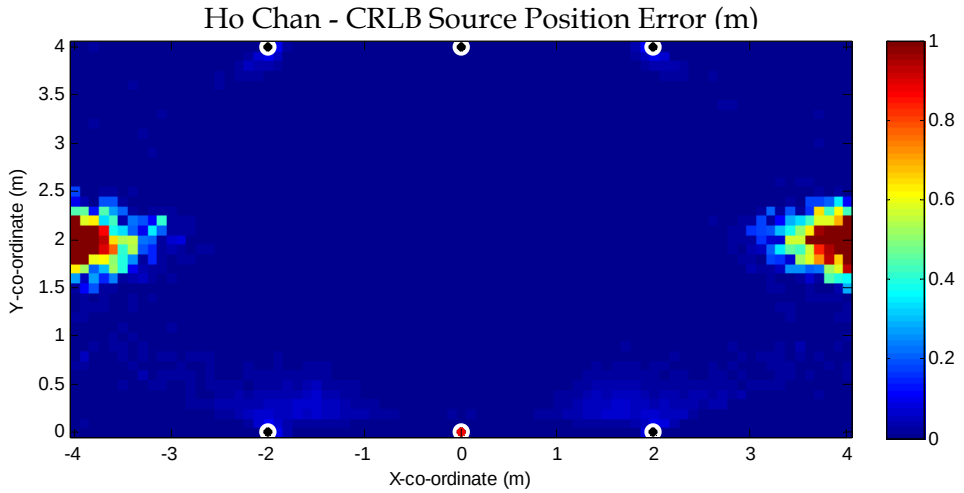

Figure 50. Plot of CRLB subtracted from Ho-Chan standard deviations; room size = 8x4 m; noise variance = 0.01 m²; reference sensor in the middle.

Figures 51-60 show the standard deviation plots of Set 4.

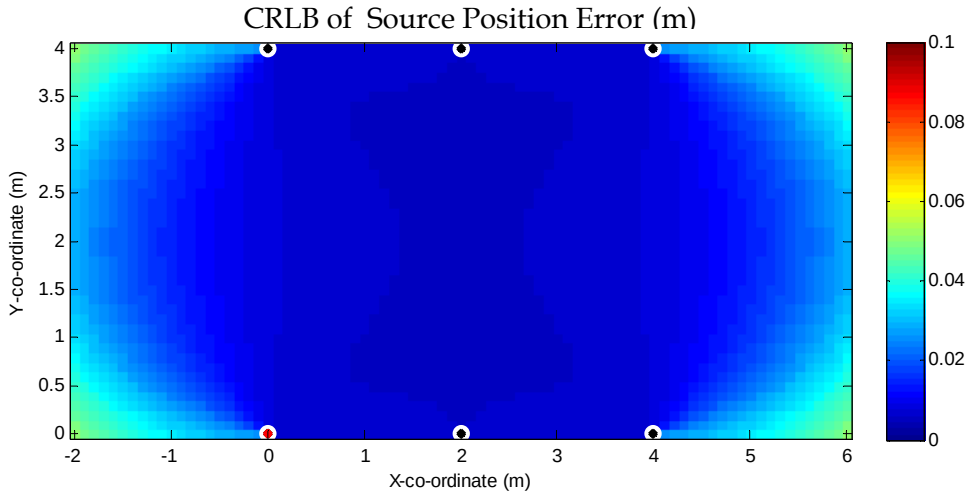

Figure 51. Standard deviation plot of CRLB of TDOA distance error; room size = 8x4 m; noise variance = 0.0001 m²; reference sensor on the side.

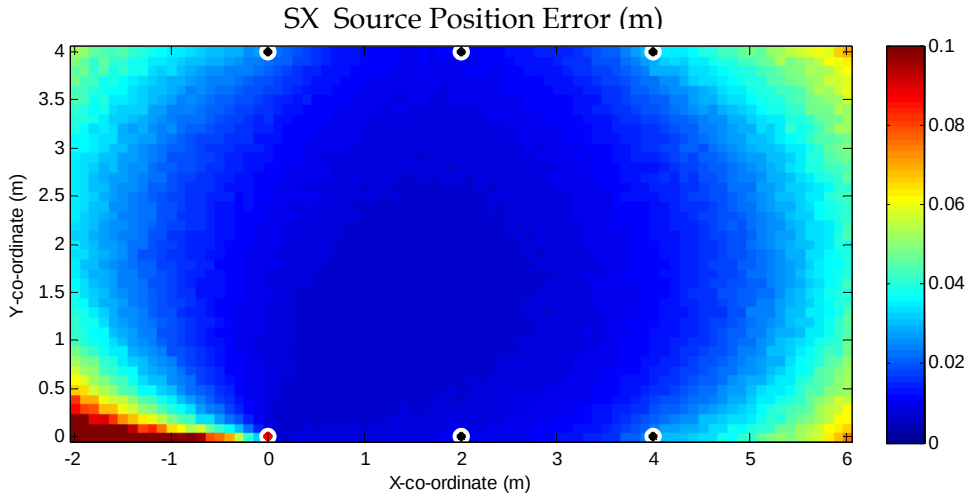

Figure 52. Standard deviation plot of SX method over 500 samples; room size = 8x4 m; noise variance = 0.0001 m²; reference sensor on the side.

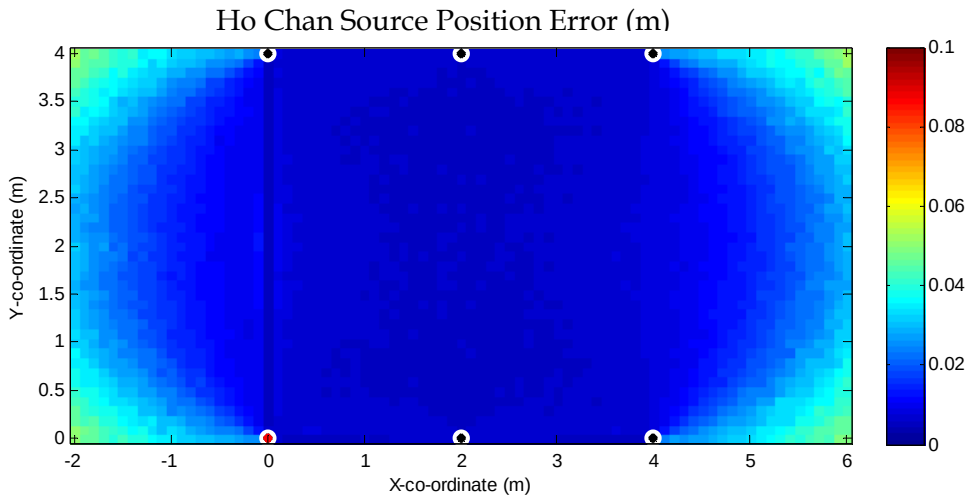

Figure 53. Standard deviation plot of Ho-Chan method over 500 samples; room size = 8x4 m; noise variance = 0.0001 m²; reference sensor on the side.

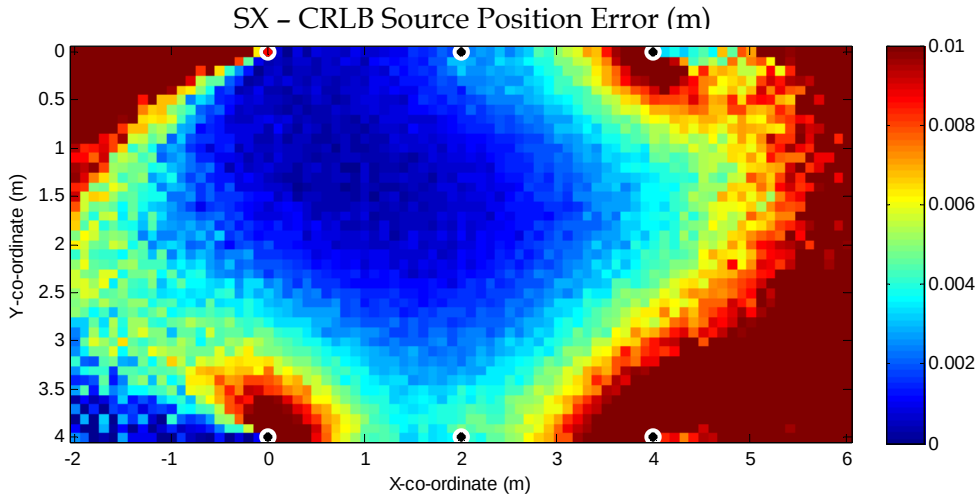

Figure 54. Plot of CRLB subtracted from SX standard deviations; room size = 8x4 m; noise variance = 0.0001 m²; reference sensor on the side.

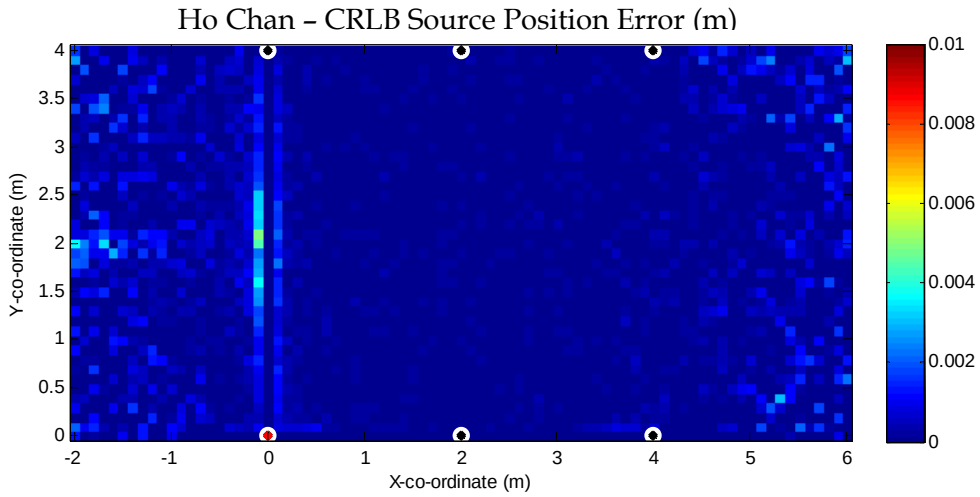

Figure 55. Plot of CRLB subtracted from Ho-Chan standard deviations; room size = 8x4 m; noise variance = 0.0001 m²; reference sensor on the side.

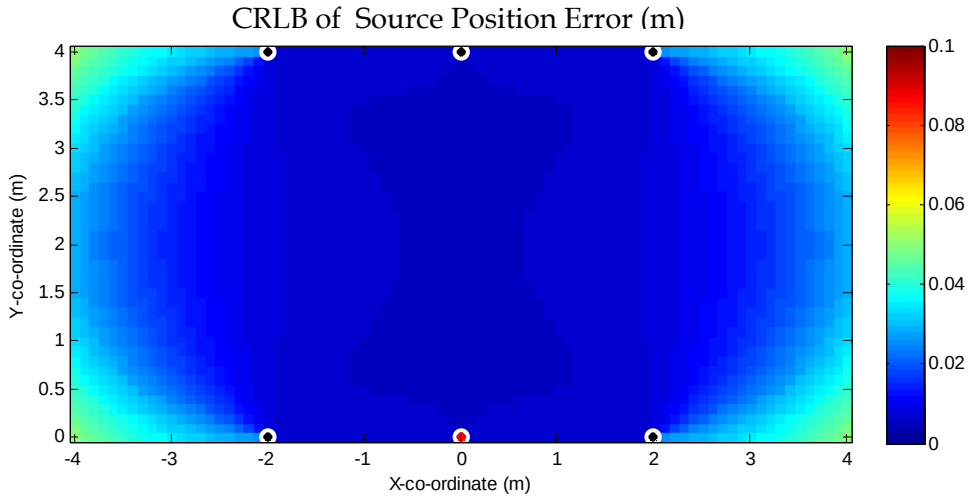

Figure 56. Standard deviation plot of CRLB of TDOA distance error; room size = 8x4 m; noise variance = 0.0001 m²; reference sensor in the middle.

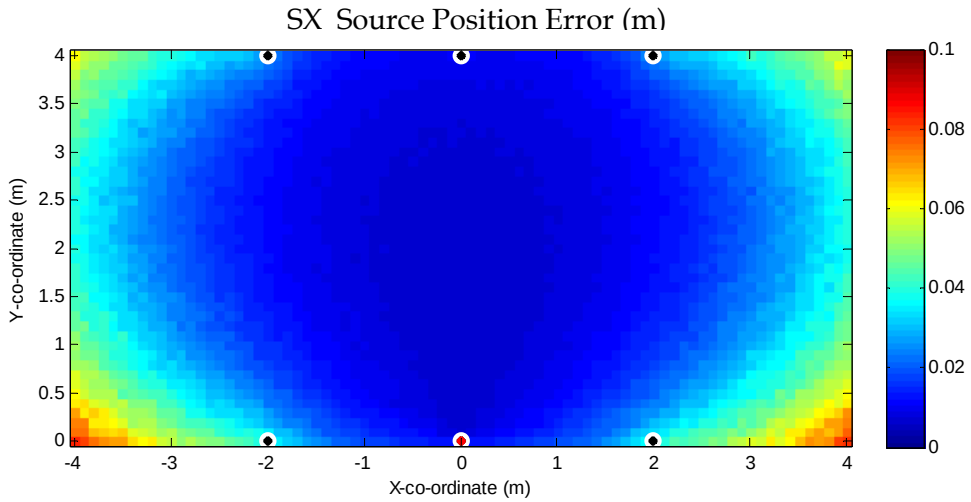

Figure 57. Standard deviation plot of SX method over 500 samples; room size = 8x4 m; noise variance = 0.0001 m²; reference sensor in the middle.

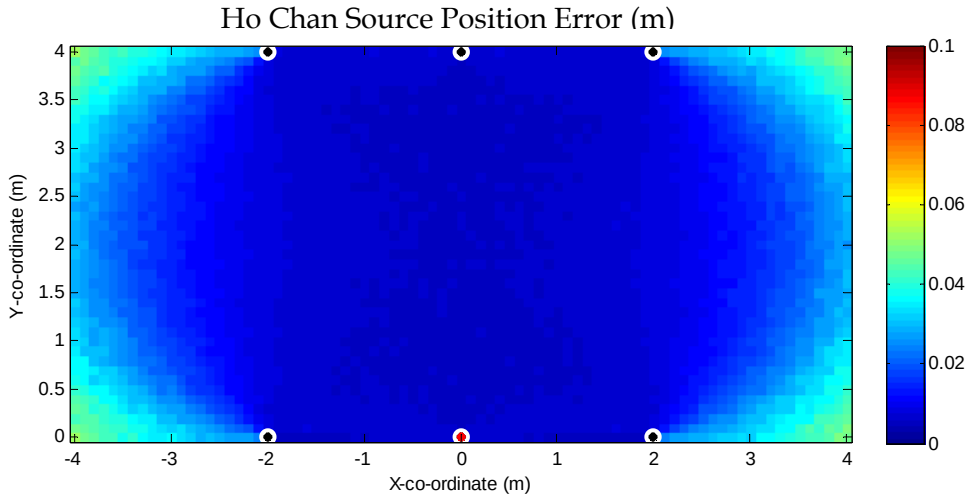

Figure 58. Standard deviation plot of Ho-Chan method over 500 samples; room size = 8x4 m; noise variance = 0.0001 m²; reference sensor in the middle.

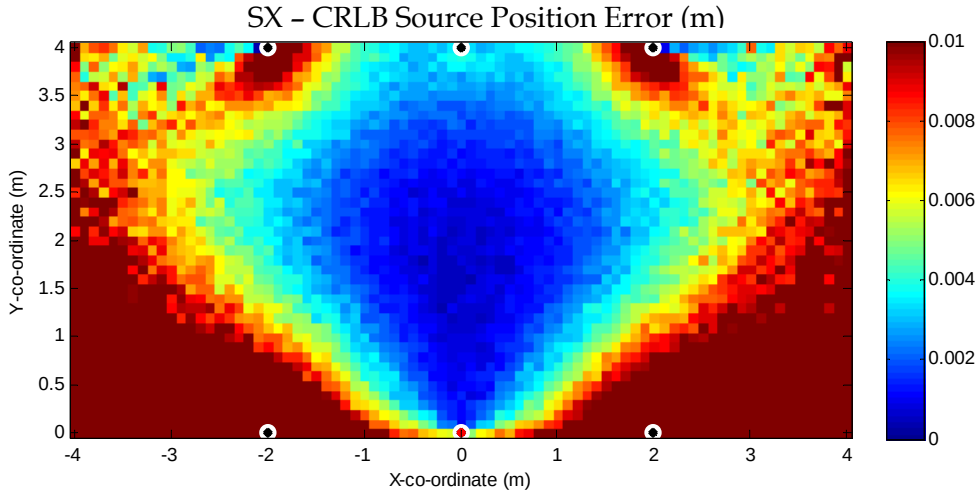

Figure 59. Plot of CRLB subtracted from SX standard deviations; room size = 8x4 m; noise variance = 0.0001 m²; reference sensor in the middle.

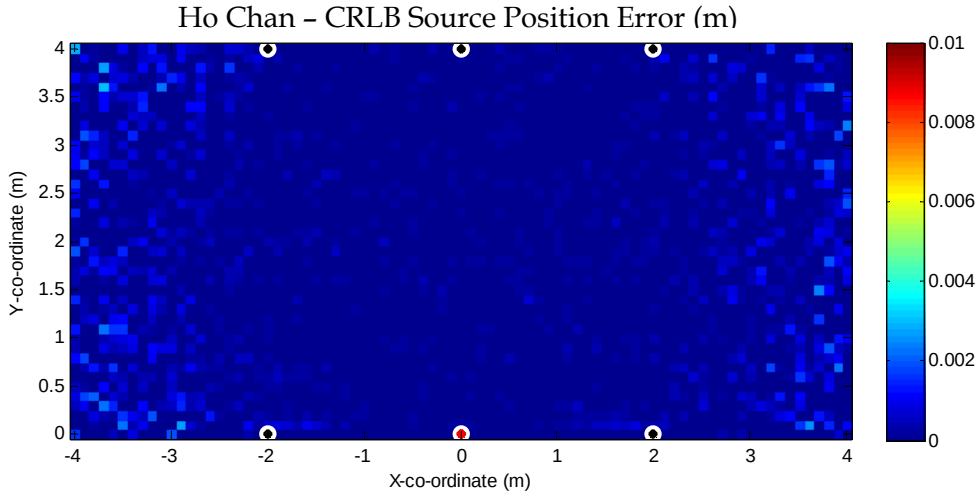

Figure 60. Plot of CRLB subtracted from Ho-Chan standard deviations; room size = 8x4 m; noise variance = 0.0001 m²; reference sensor in the middle.

Figures 61-65 are have the same room size and noise variance as Set 4. The difference is that the microphones are randomly placed along the walls.

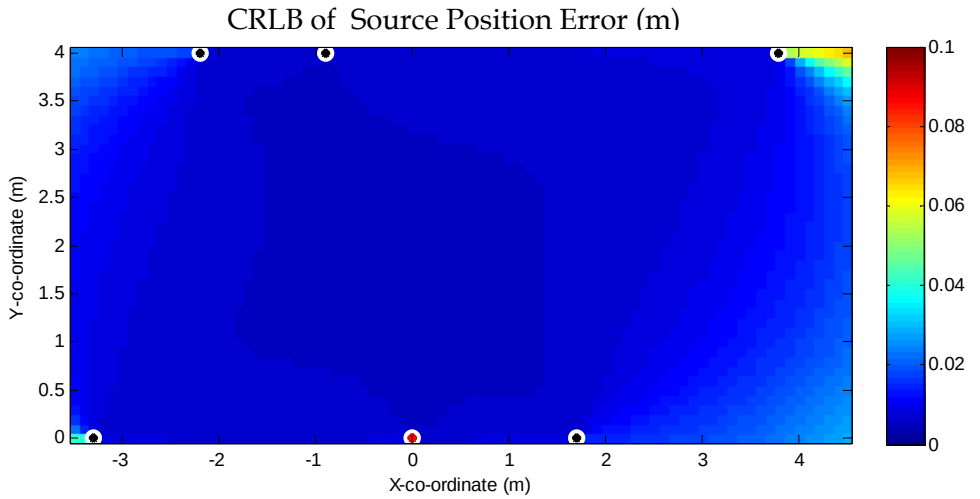

Figure 61. Standard deviation plot of CRLB of TDOA distance error; room size = 8x4 m; noise variance = 0.0001 m²; random microphone array.

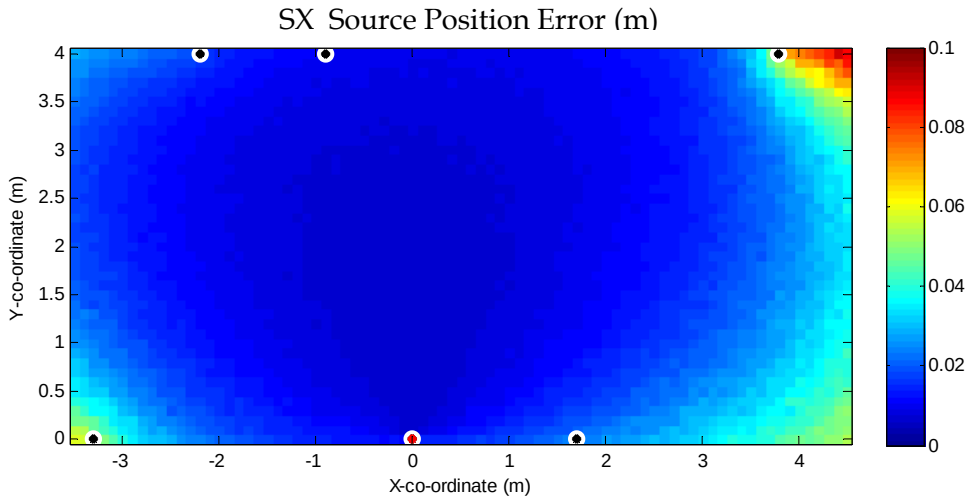

Figure 62. Standard deviation plot of SX method over 500 samples; room size = 8x4 m; noise variance = 0.0001 m²; random microphone array.

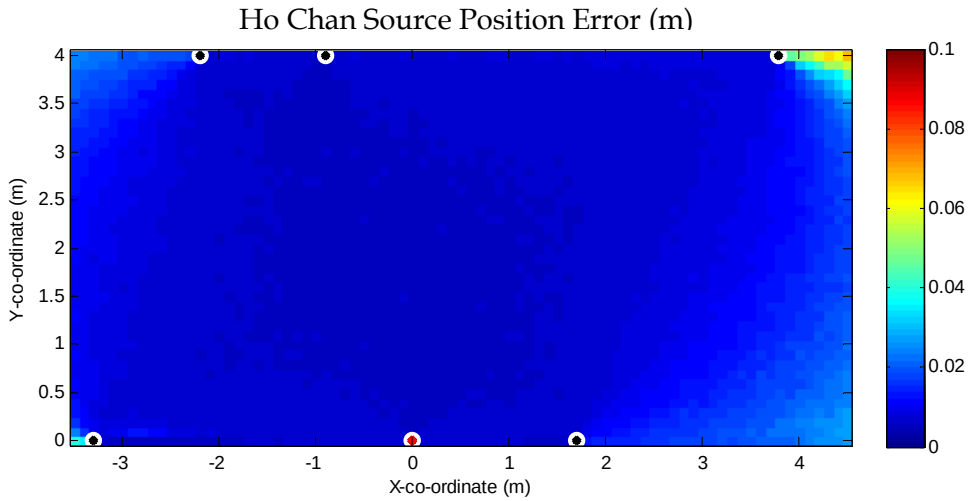

Figure 63. Standard deviation plot of Ho-Chan method over 500 samples; room size = 8x4 m; noise variance = 0.0001 m²; random microphone array.

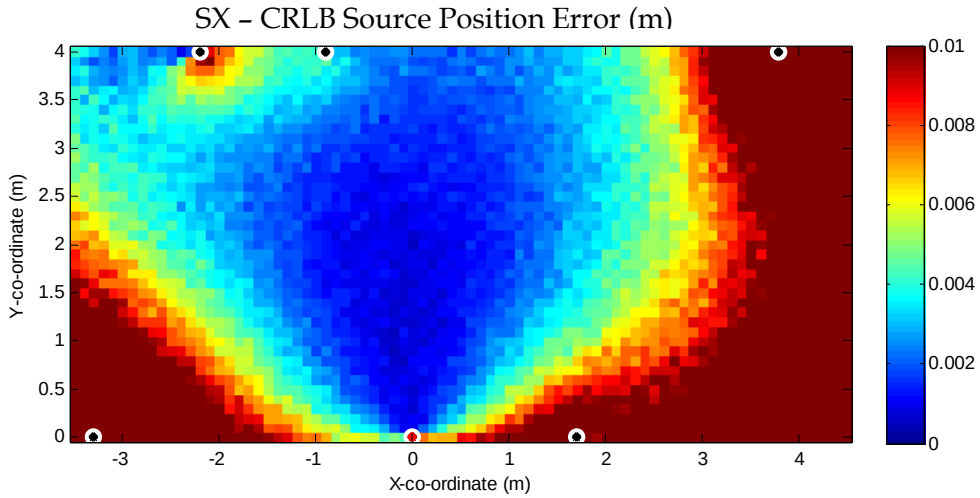

Figure 64. Plot of CRLB subtracted from SX standard deviations; room size = 8x4 m; noise variance = 0.0001 m²; random microphone array.

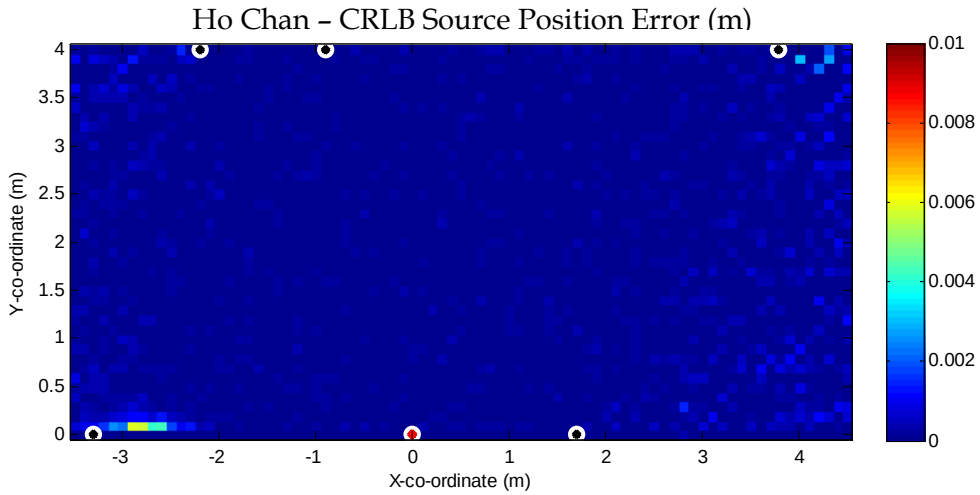

Figure 65. Plot of CRLB subtracted from Ho-Chan standard deviations; room size = 8x4 m; noise variance = 0.0001 m²; random microphone array.

Appendix C: Spectral Plots of Experiment Recordings

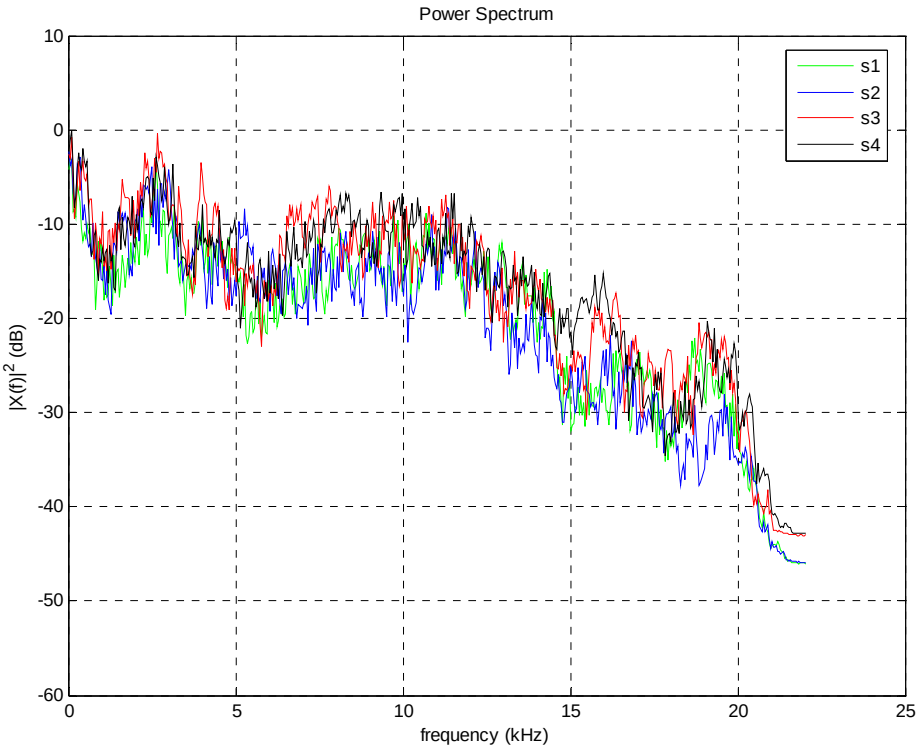

Figure 66. White noise, overload.

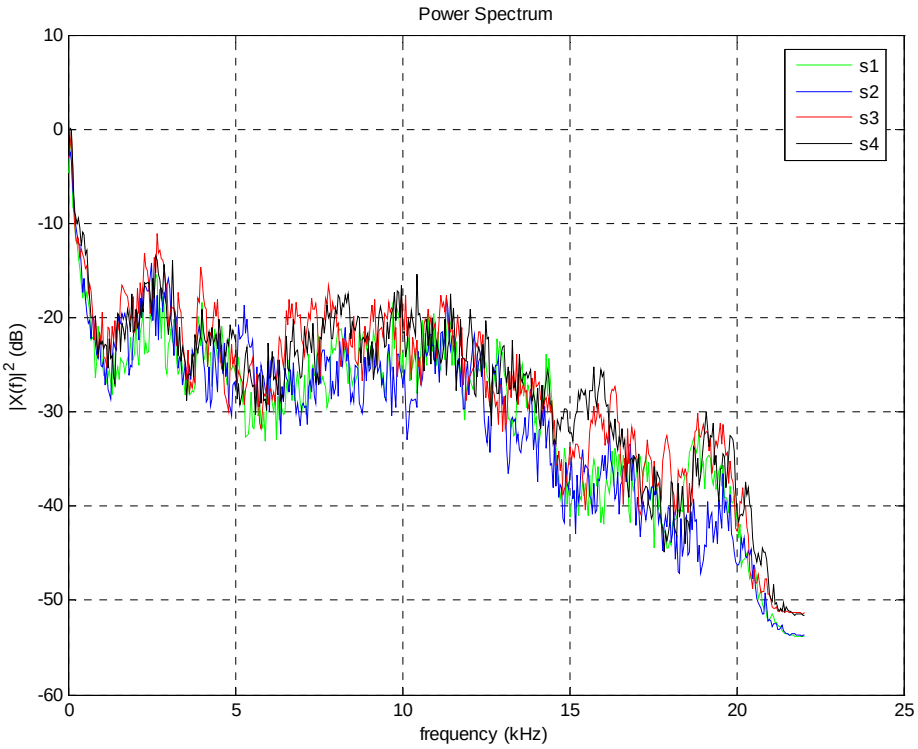

Figure 67. White noise, 90dBA.

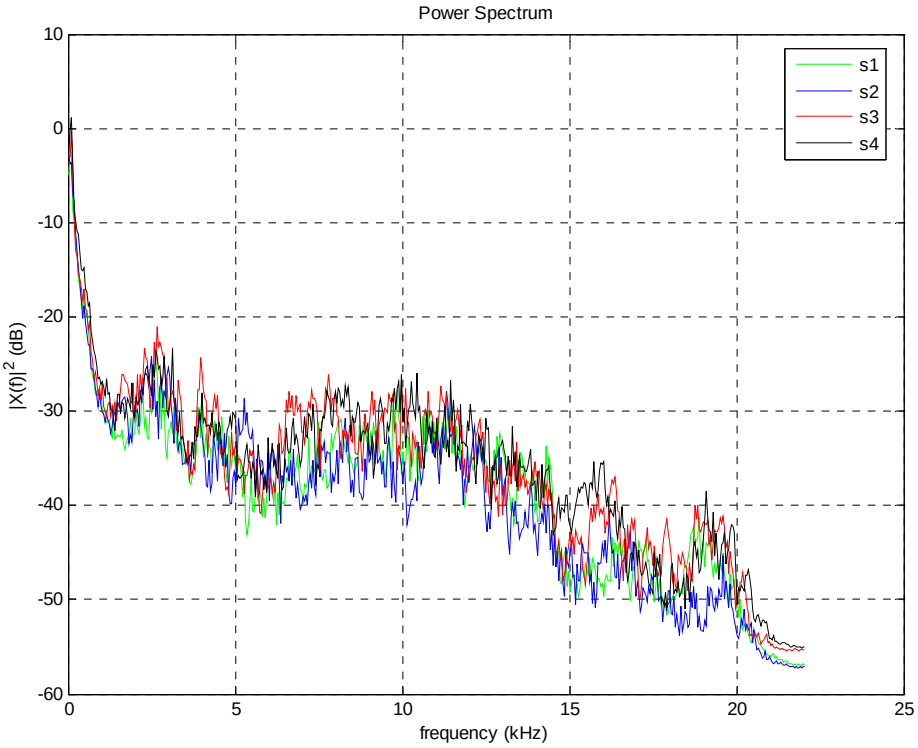

Figure 68. White noise, 80dBA.

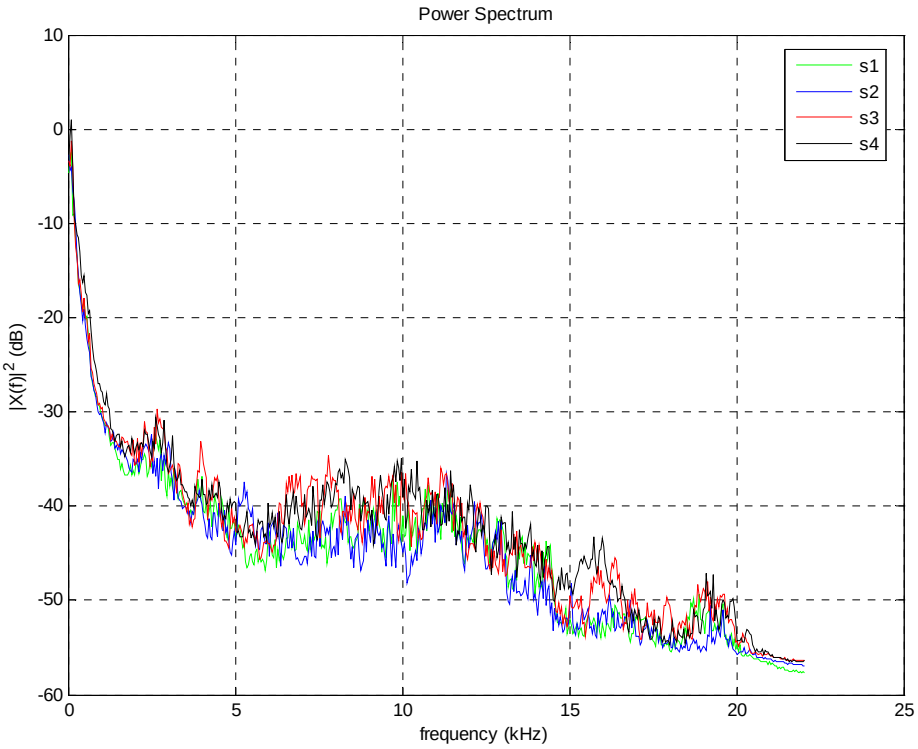

Figure 69. White noise, 70dBA.

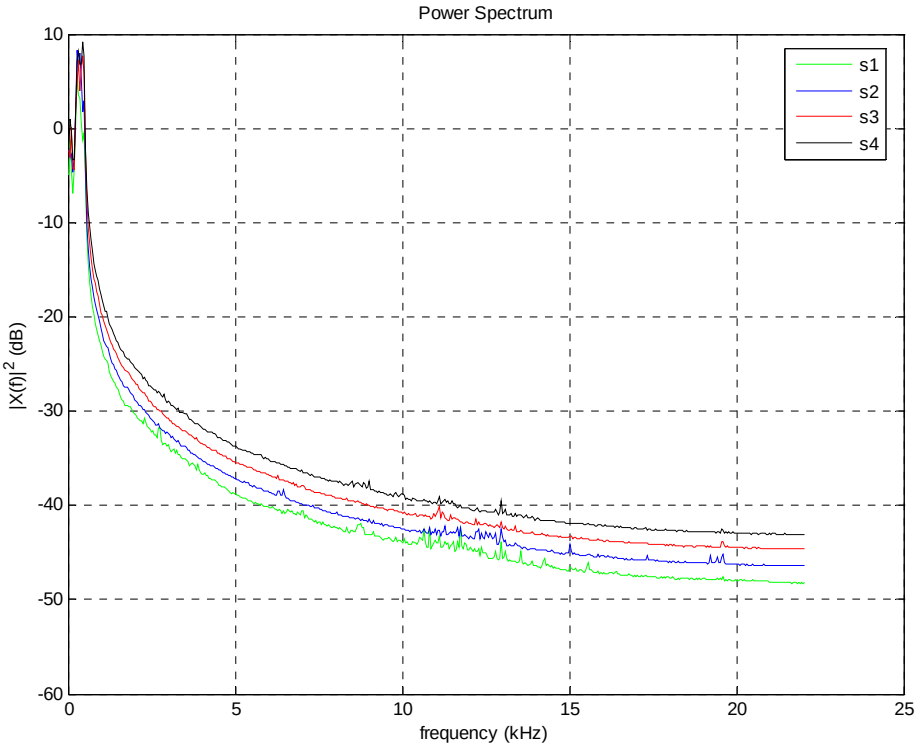

Figure 70. 50-500Hz noise, 95dBA.

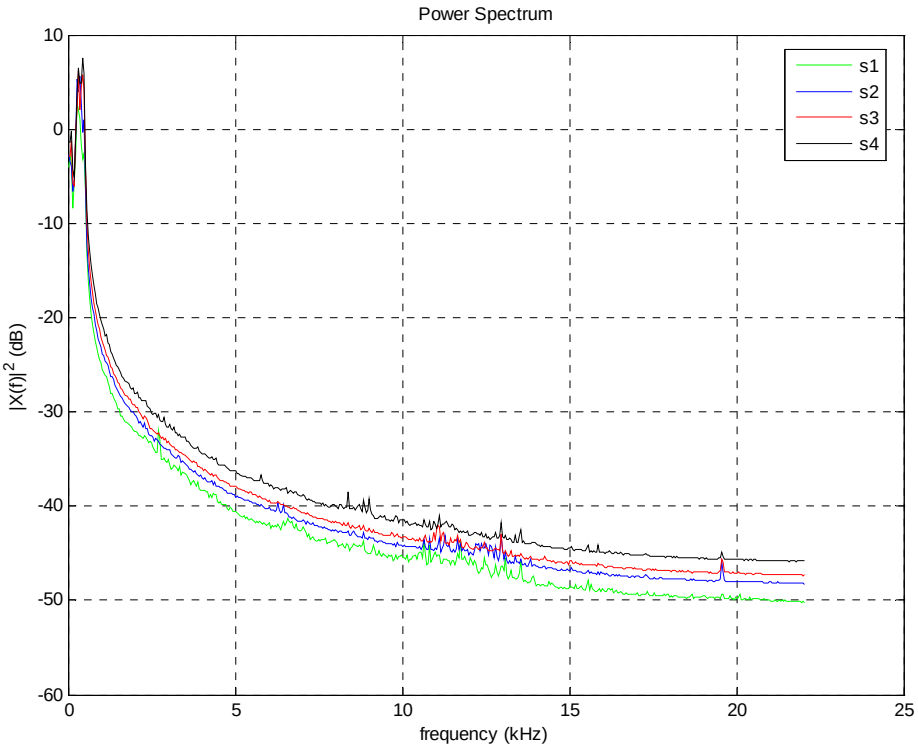

Figure 71. 50-500Hz noise, 90dBA.

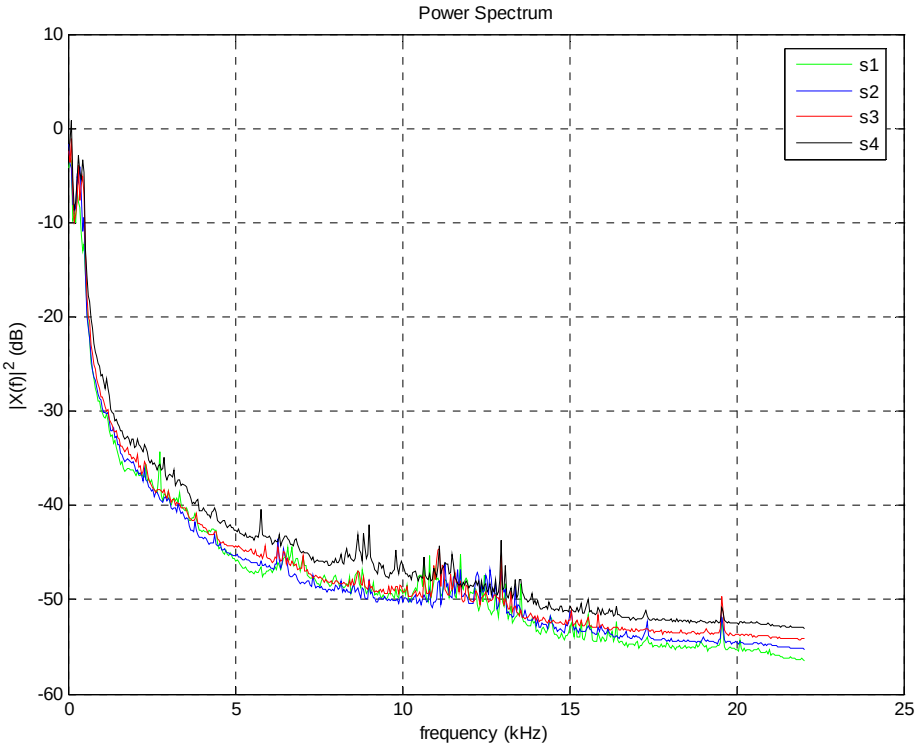

Figure 72. 50-500Hz noise, 80dBA.

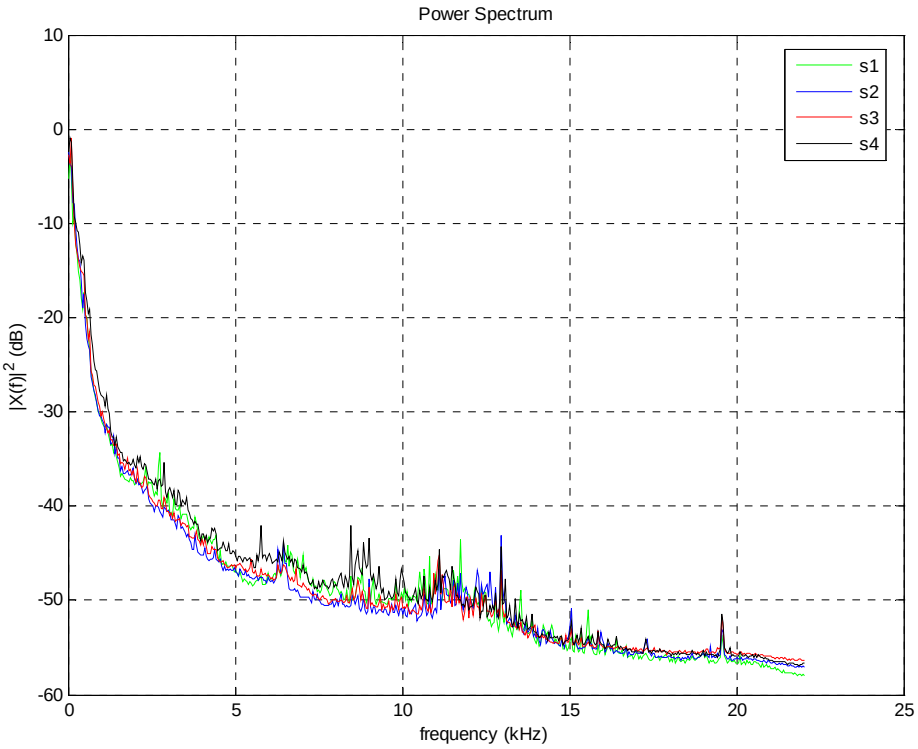

Figure 73. 50-500Hz noise, 70dBA.

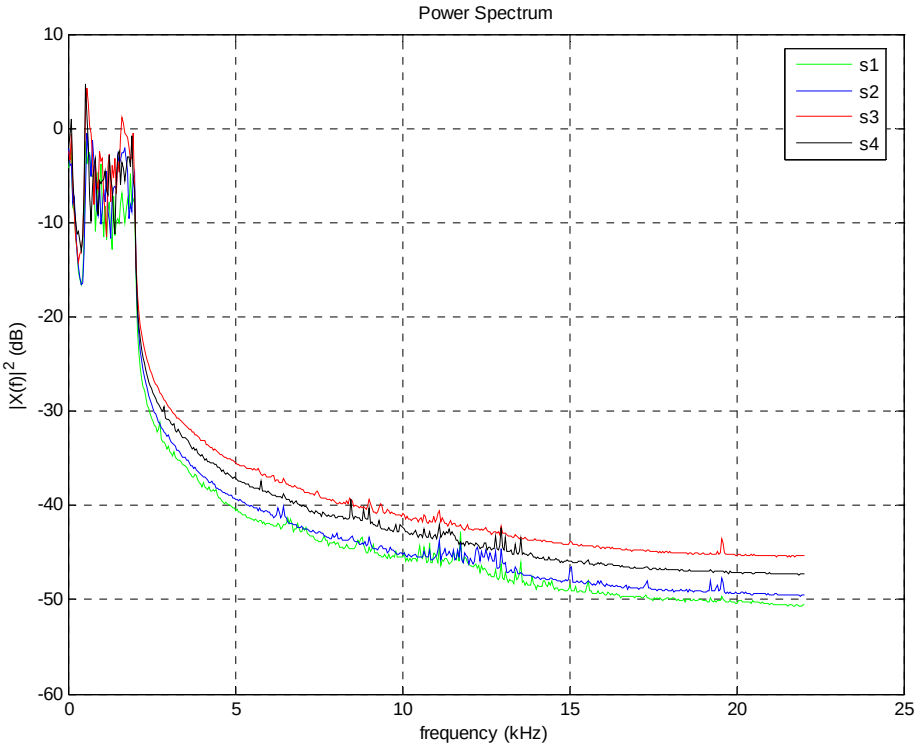

Figure 74. 500-2000Hz noise, overload.

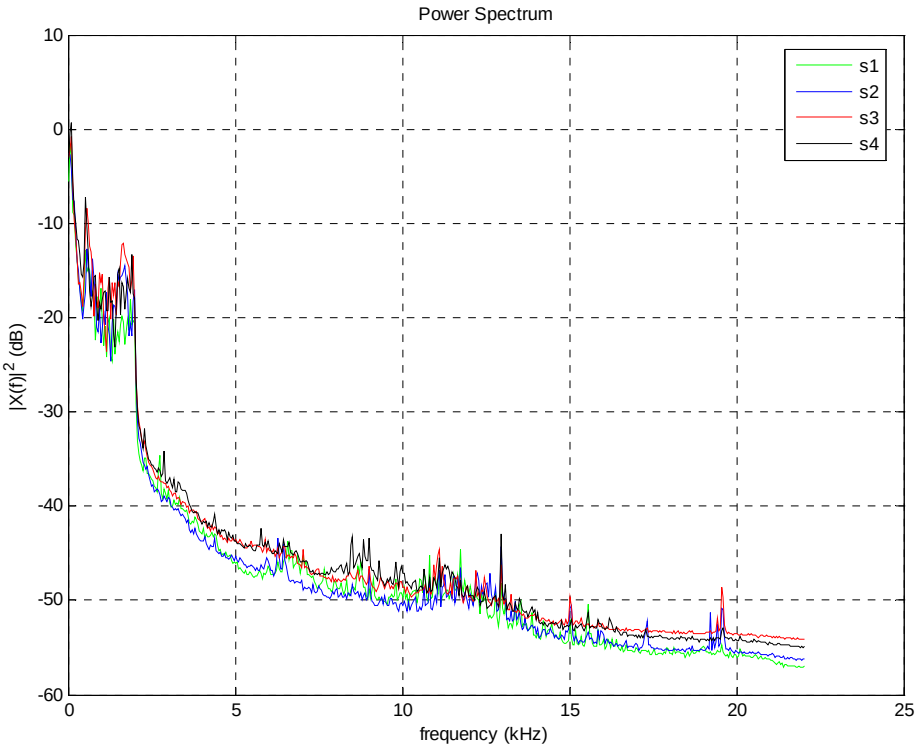

Figure 75. 500-2000Hz noise, 90dBA.

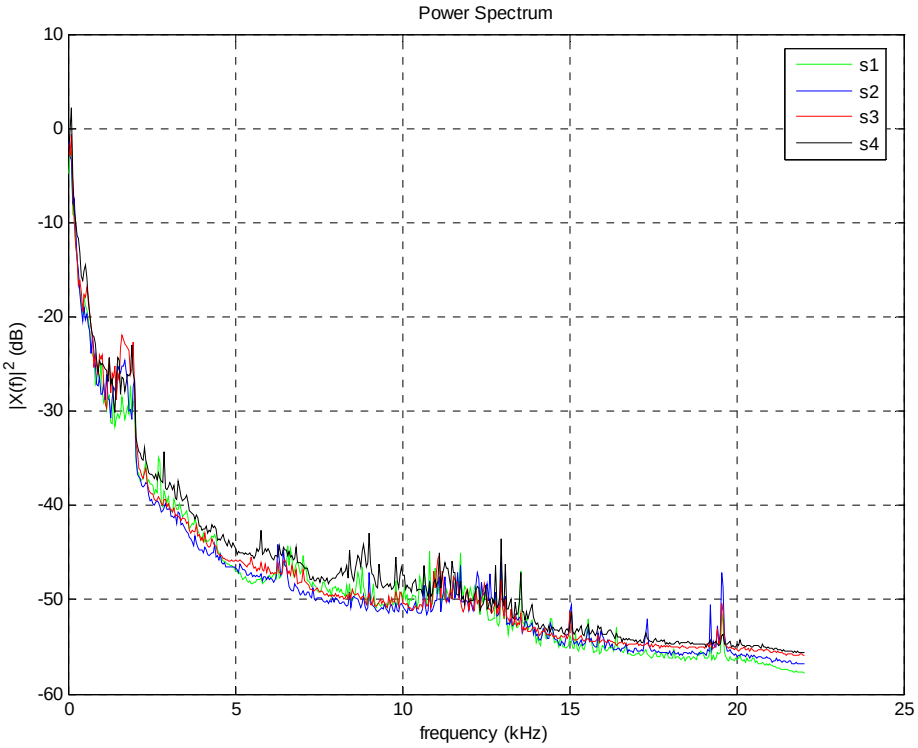

Figure 76. 500-2000Hz noise, 80dBA.

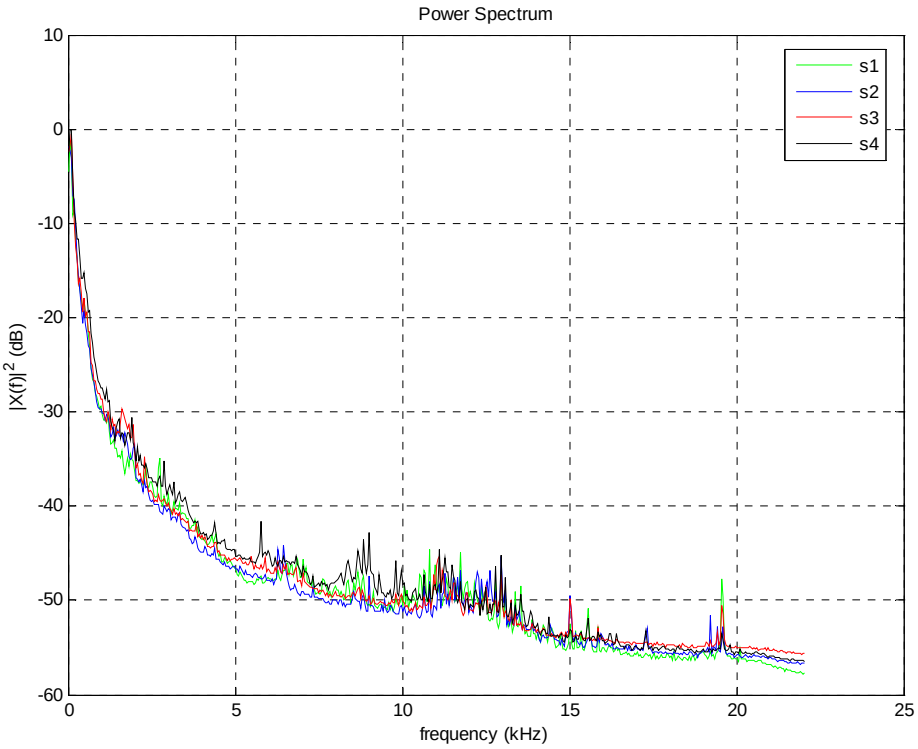

Figure 77. 500-2000Hz noise, 70dBA.

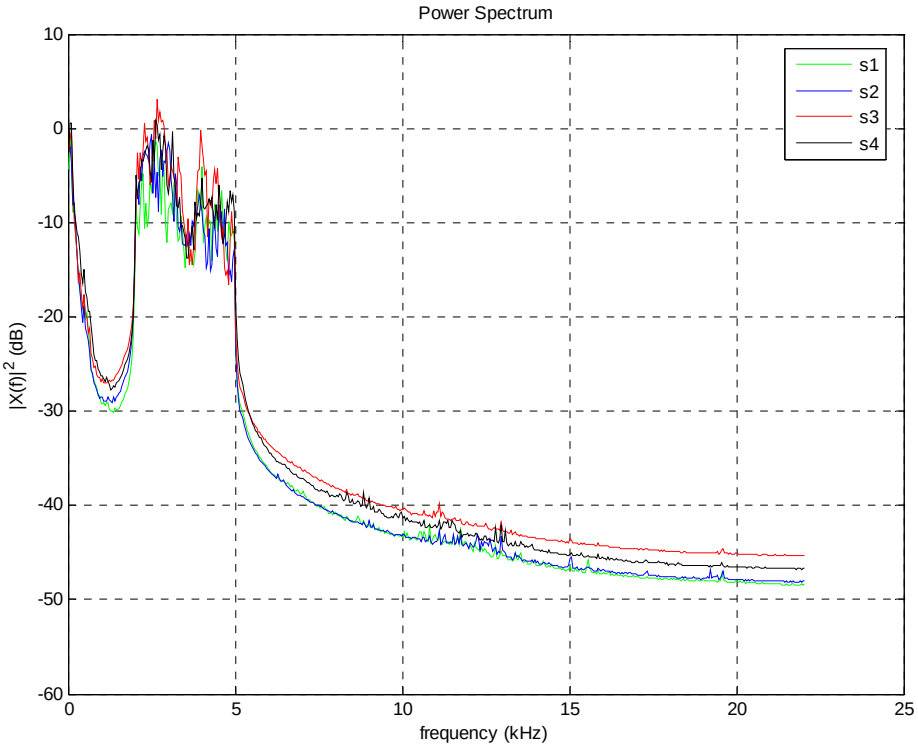

Figure 78. 2000-5000Hz noise, overload.

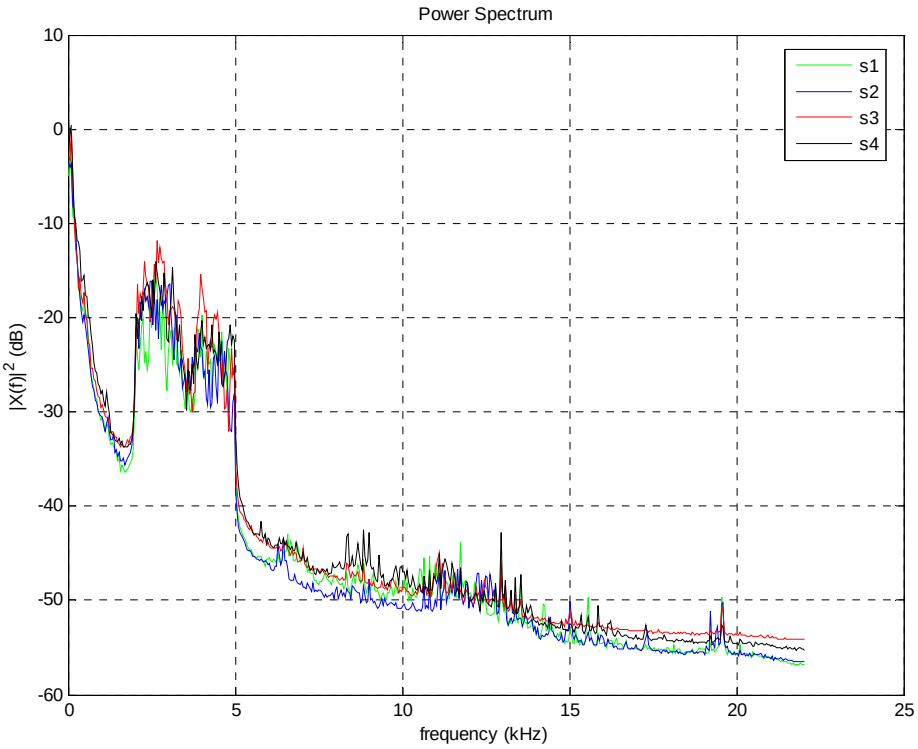

Figure 79. 2000-5000Hz noise, 90dBA.

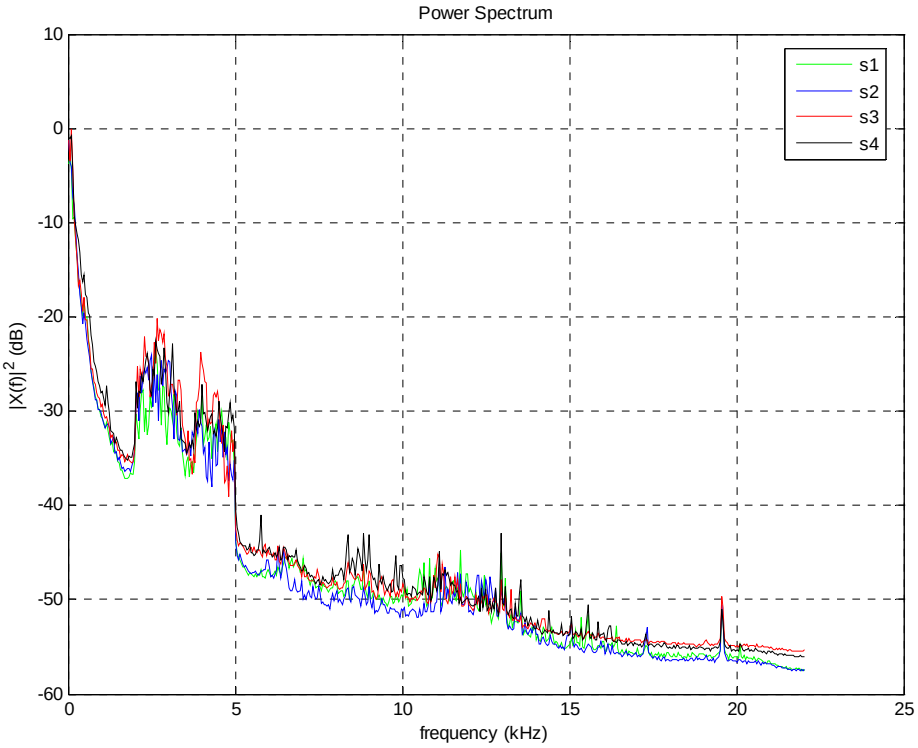

Figure 80. 2000-5000Hz noise, 80dBA.

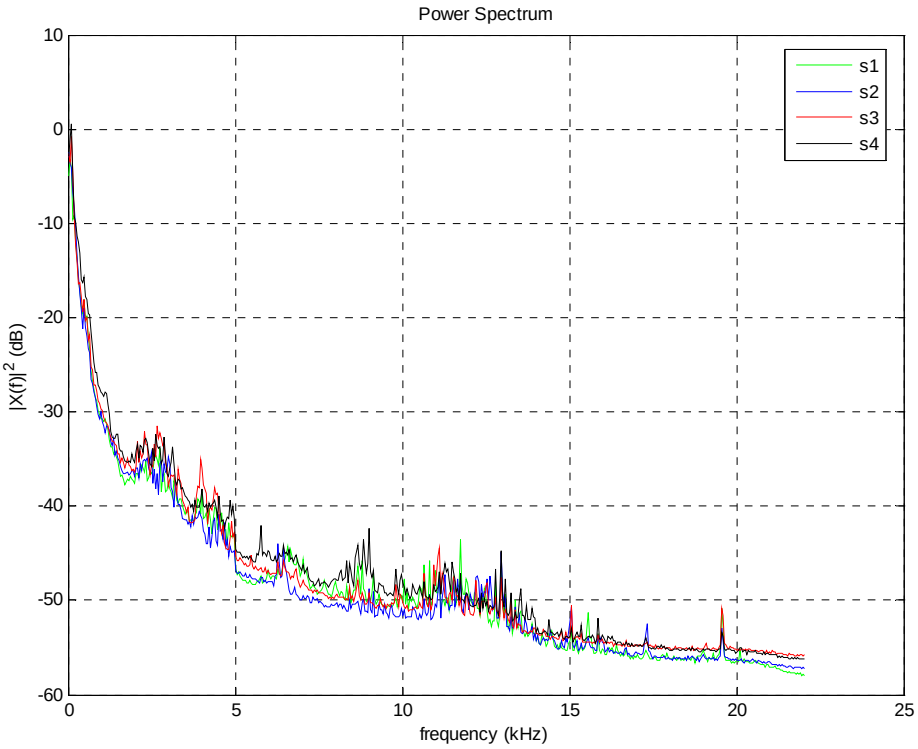

Figure 81. 2000-5000Hz noise, 70dBA.

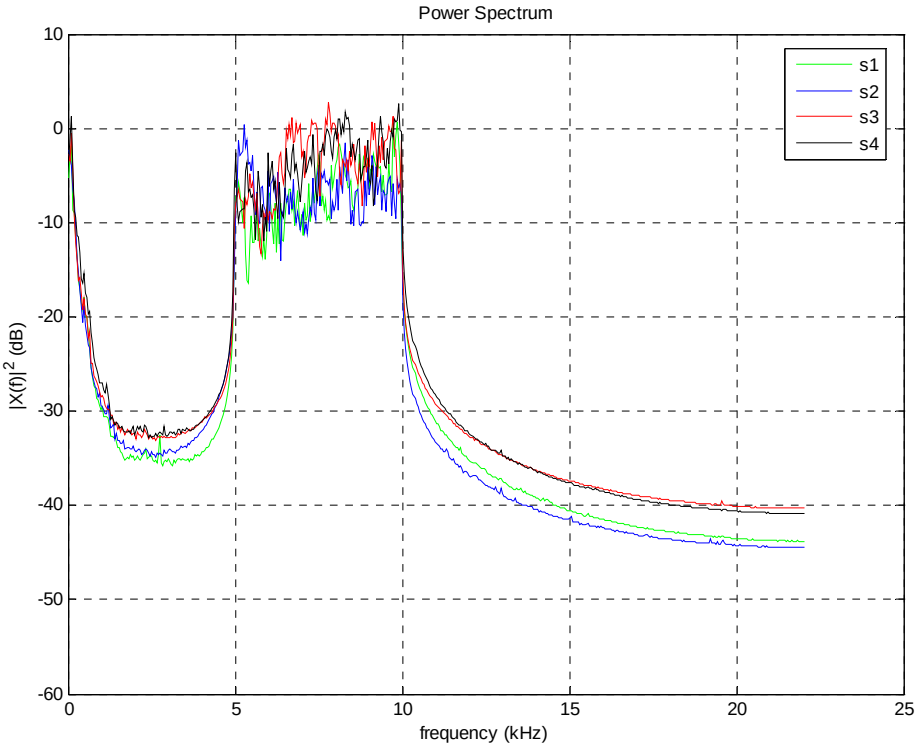

Figure 82. 5000-10000Hz noise, overload.

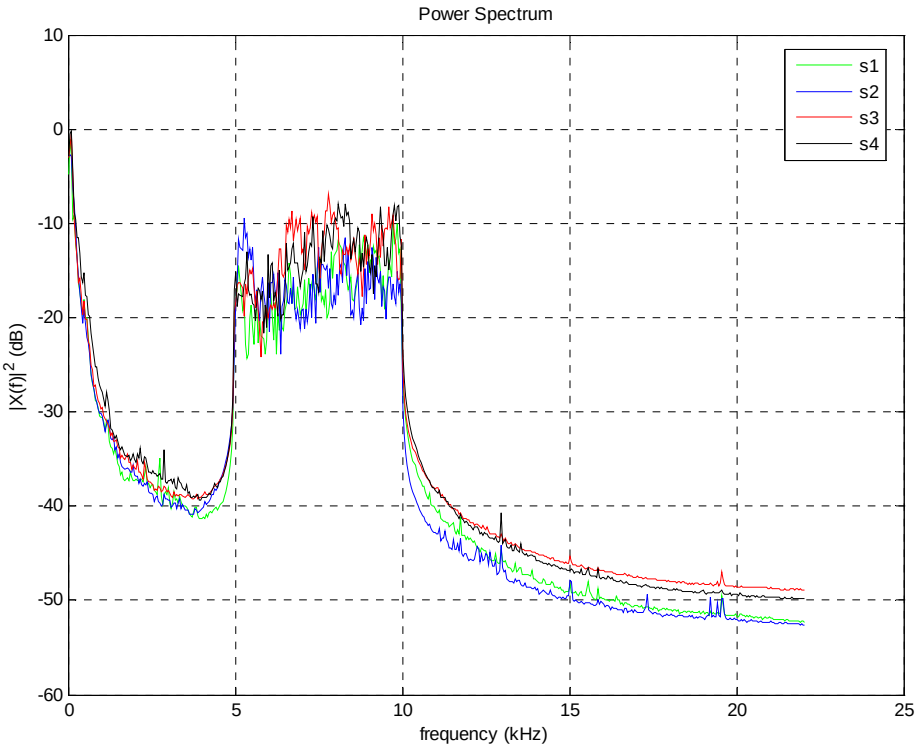

Figure 83. 5000-10000Hz noise, 90dBA.

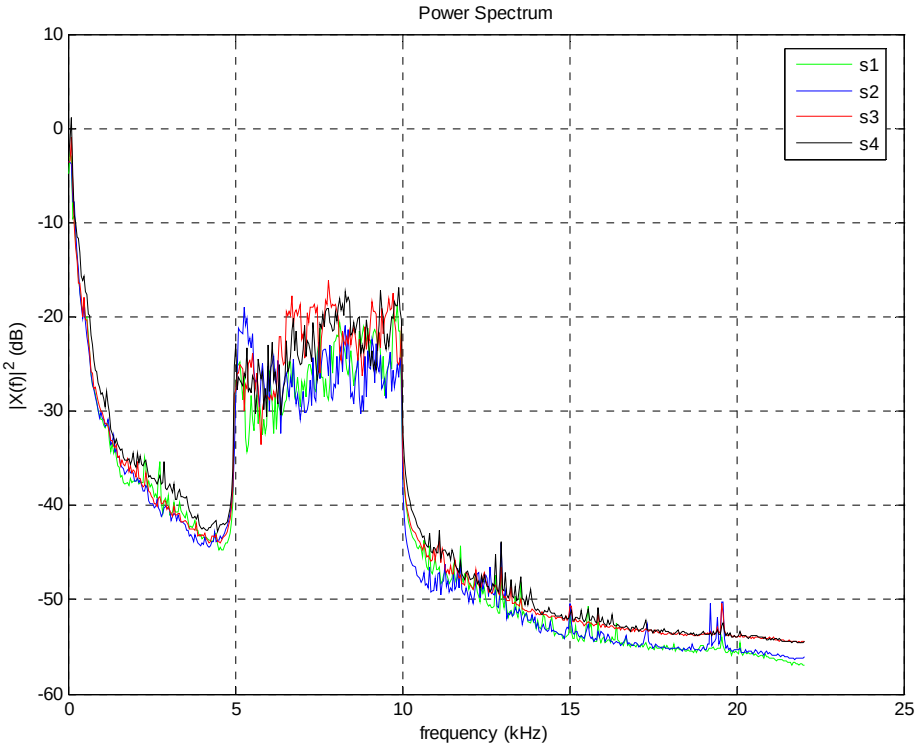

Figure 84. 5000-10000Hz noise, 80dBA.

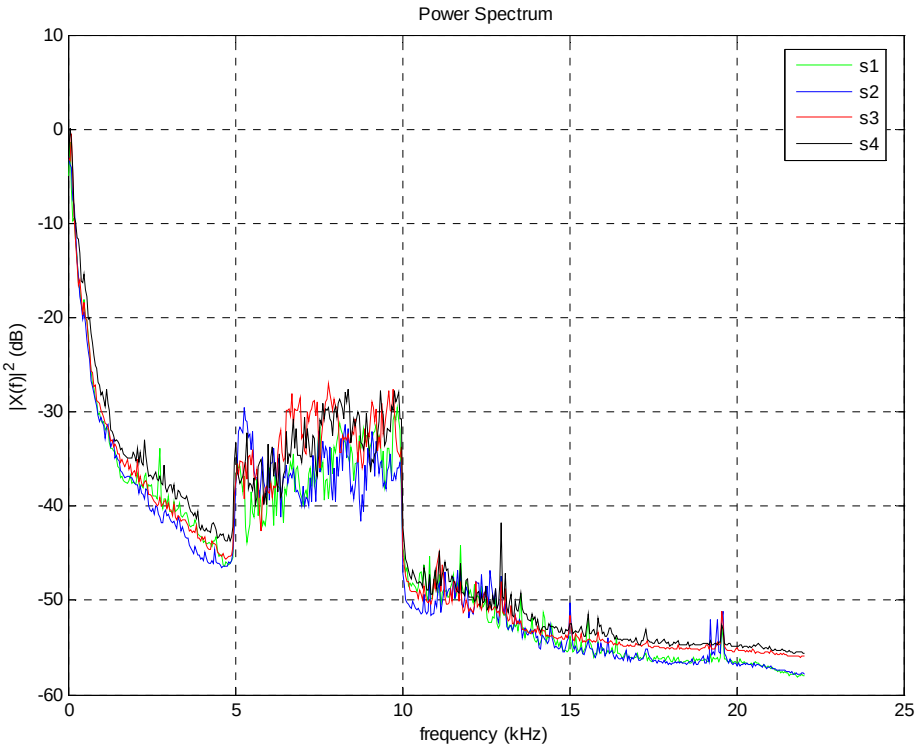

Figure 85. 5000-10000Hz noise, 70dBA.

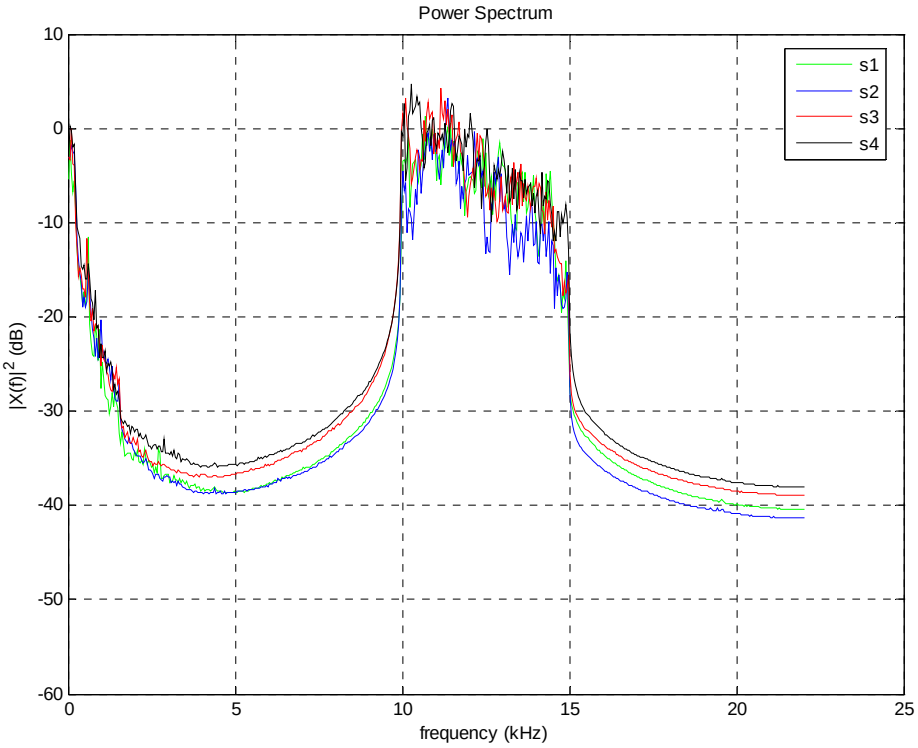

Figure 86. 10000-15000Hz noise, overload.

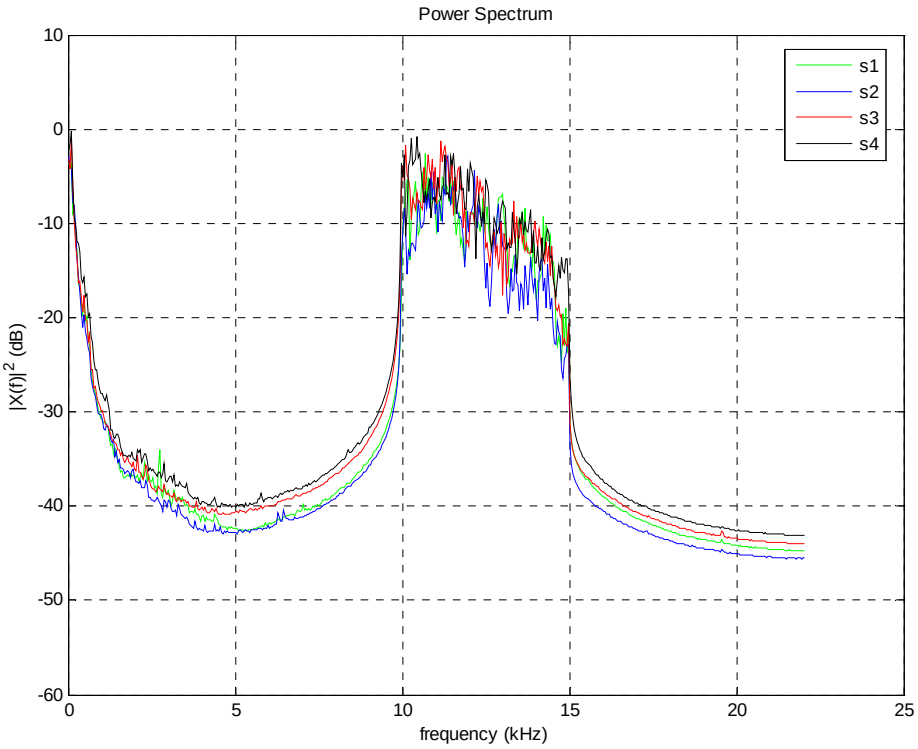

Figure 87. 10000-15000Hz noise, 90dBA.

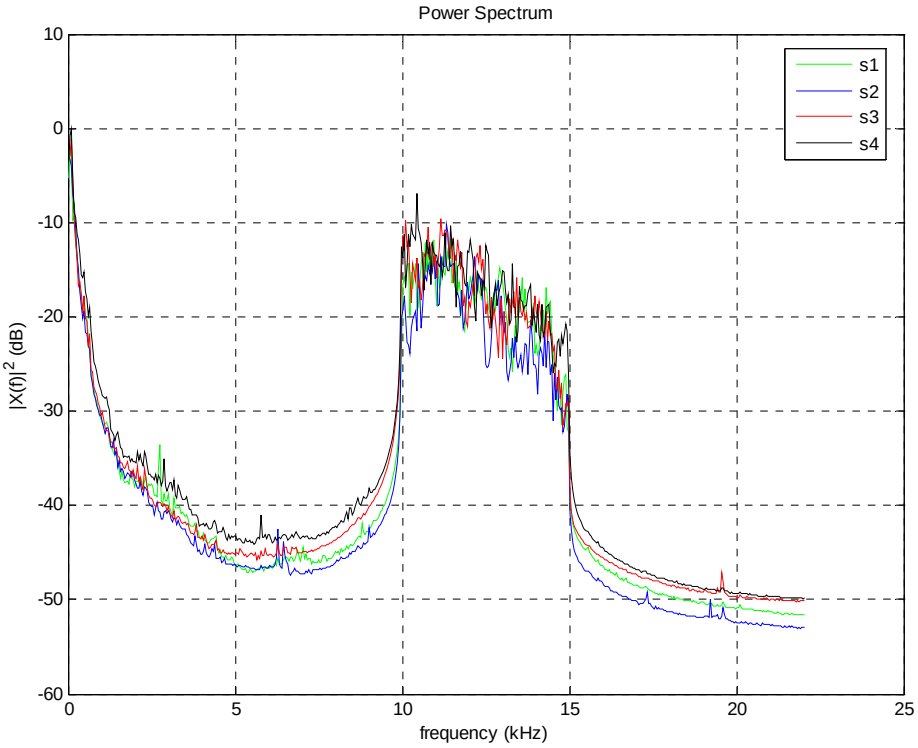

Figure 88. 10000-15000Hz noise, 80dBA.

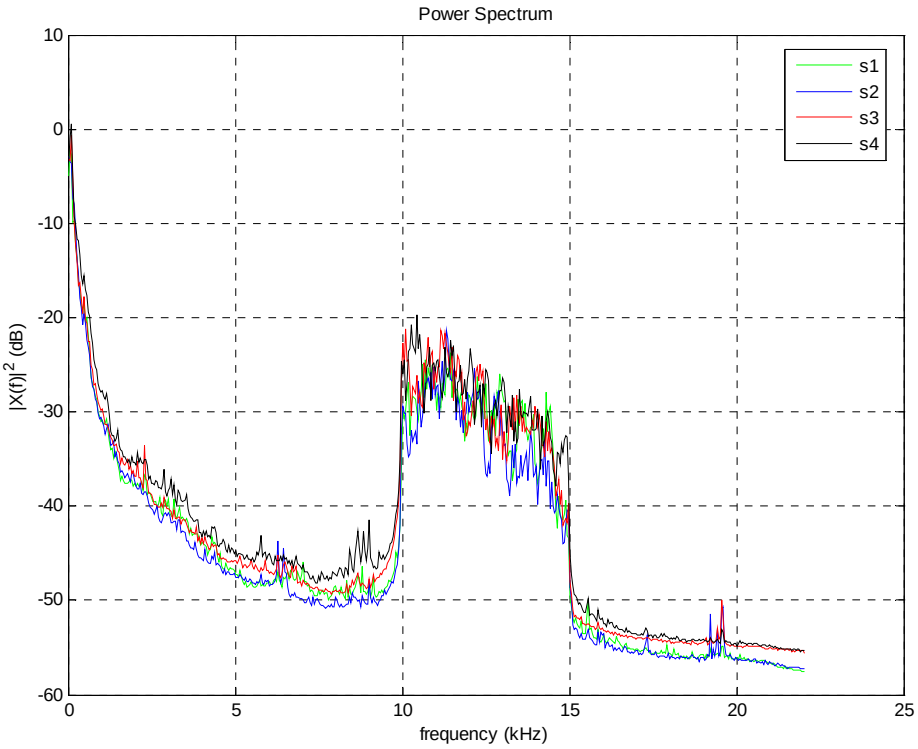

Figure 89. 10000-15000Hz noise, 70dBA.

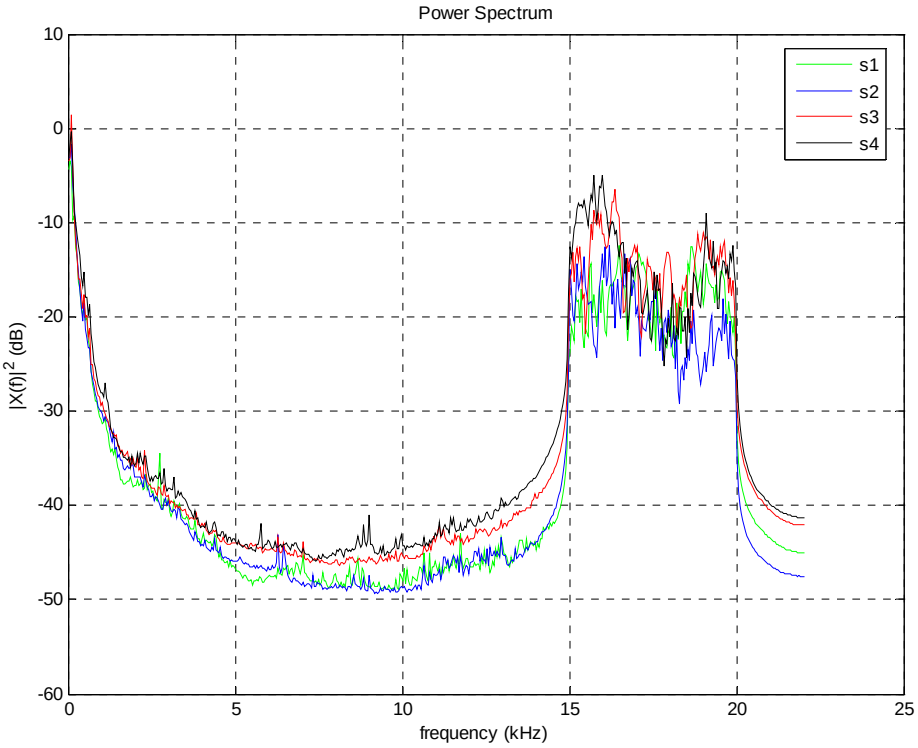

Figure 90. 15000-20000Hz noise, 75dBA.

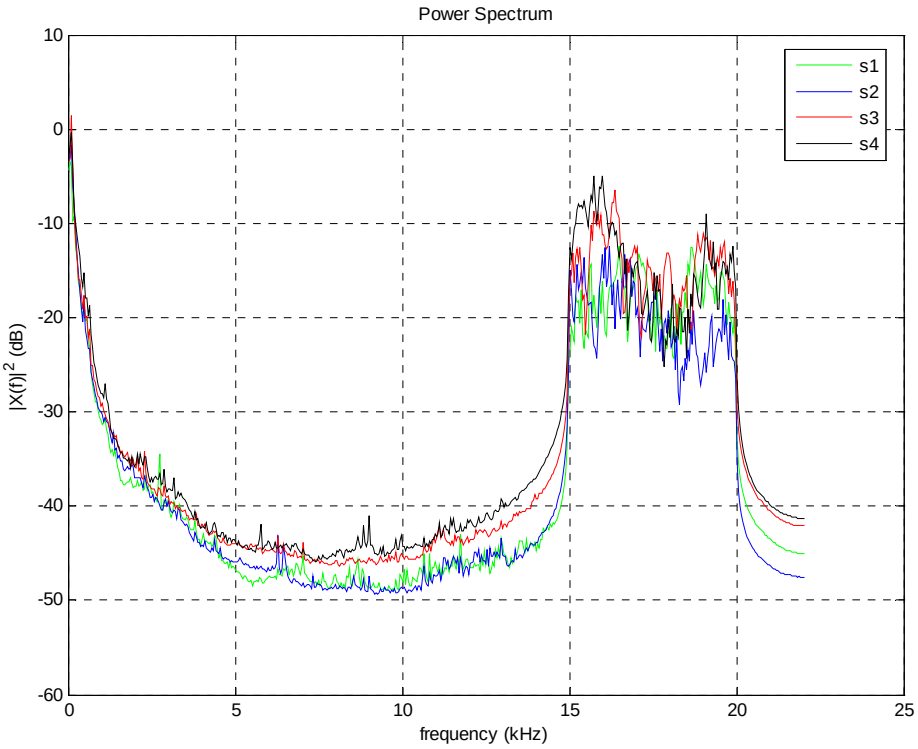

Figure 91. 15000-20000Hz noise, 70dBA.

Appendix D: Plots of Time-delay Estimates

Figures 92-99 show time-delay estimate plots using the three time-delay estimation methods being cross-correlation, GCC-PHAT and ED methods.

The true time-delays were calculated in metres from the source and microphone coordinates and are drawn as blue lines. The estimated time-delays were calculated in samples and converted to distance in metres by dividing by the sampling rate (44100 Hz) and multiplying by the speed of sound in air at 20°C (345.1 ms^{-1}). These estimates are plotted in red dots. The 'index' value on the y-axis of plots represent which sample block the estimate came from after the large samples were broken into smaller samples, e.g. an estimate with index of 10 means that estimate came from the 10^{th} small sample block of the large sample (see Section 7 for explanation).

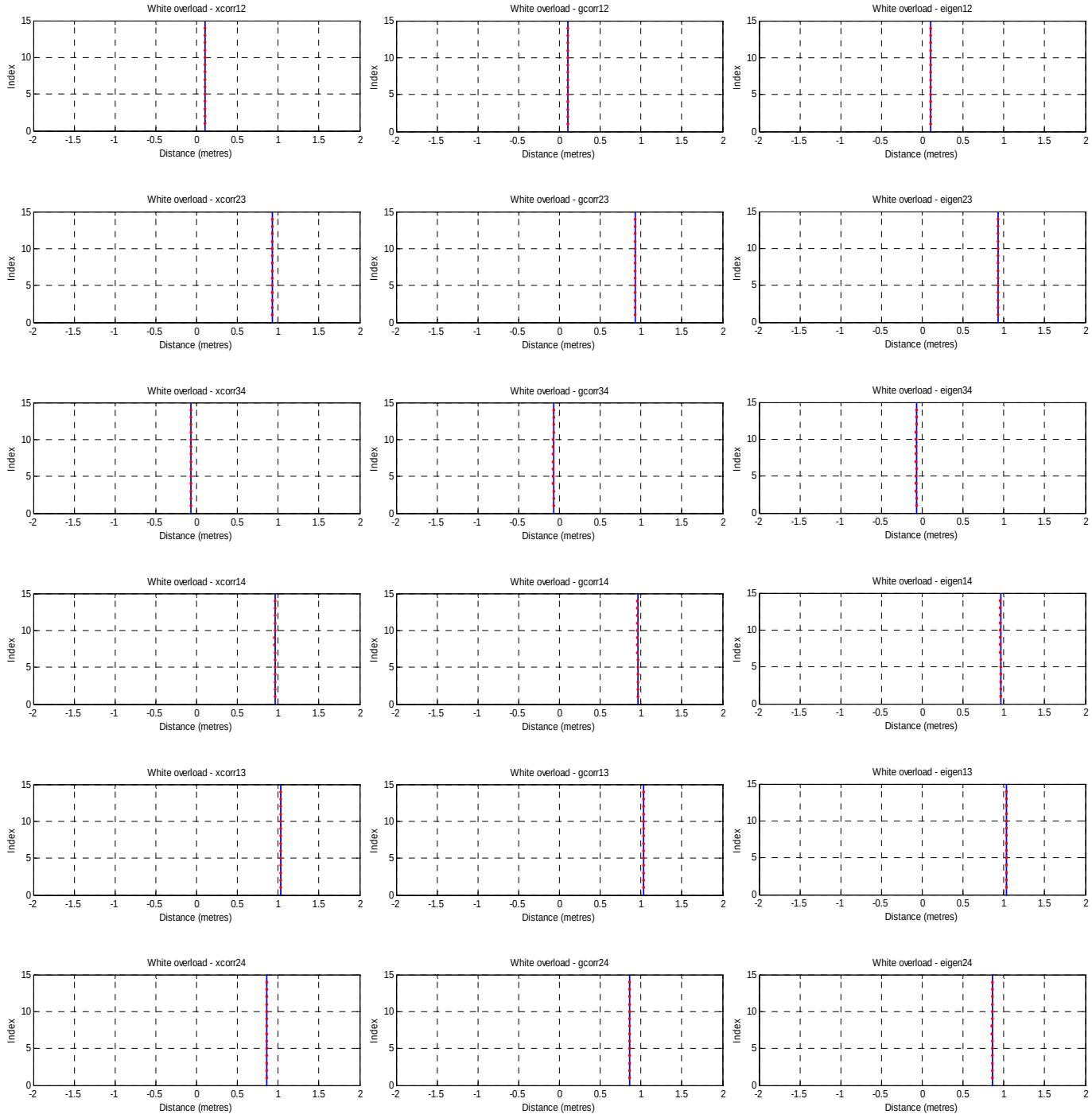

Figure 92. Time-delay estimate plots using cross-correlation, GCC-PHAT and ED method for 'overload' white noise. The red plots represent the estimates; the blue lines are the true time-delays.

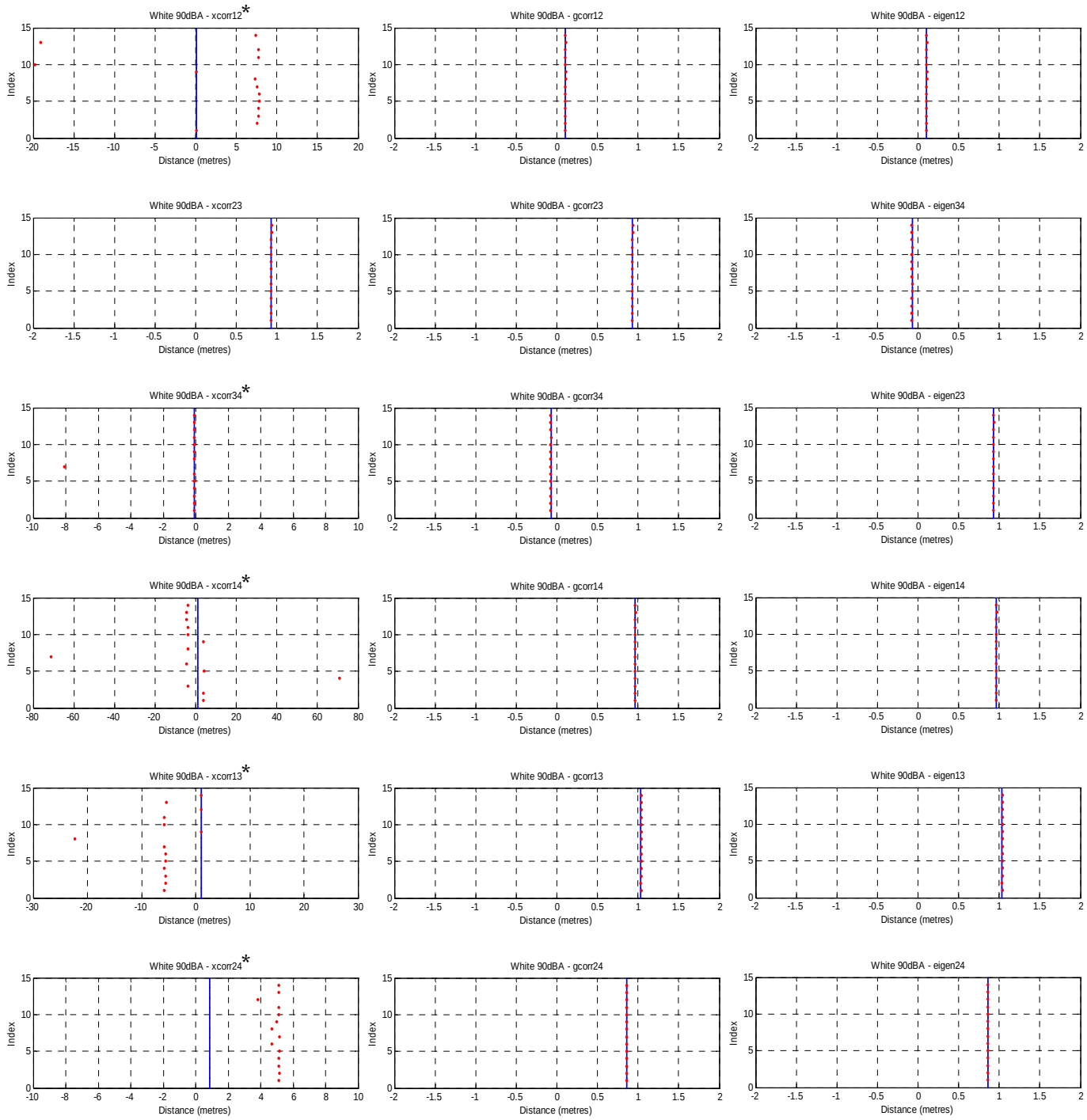

*Figure 93. Time-delay estimate plots using cross-correlation, GCC-PHAT and ED method for '90dBA' white noise. The red plots represent the estimates; the blue lines are the true time-delays; * indicates that the x-axis of that plot has a different scale.*

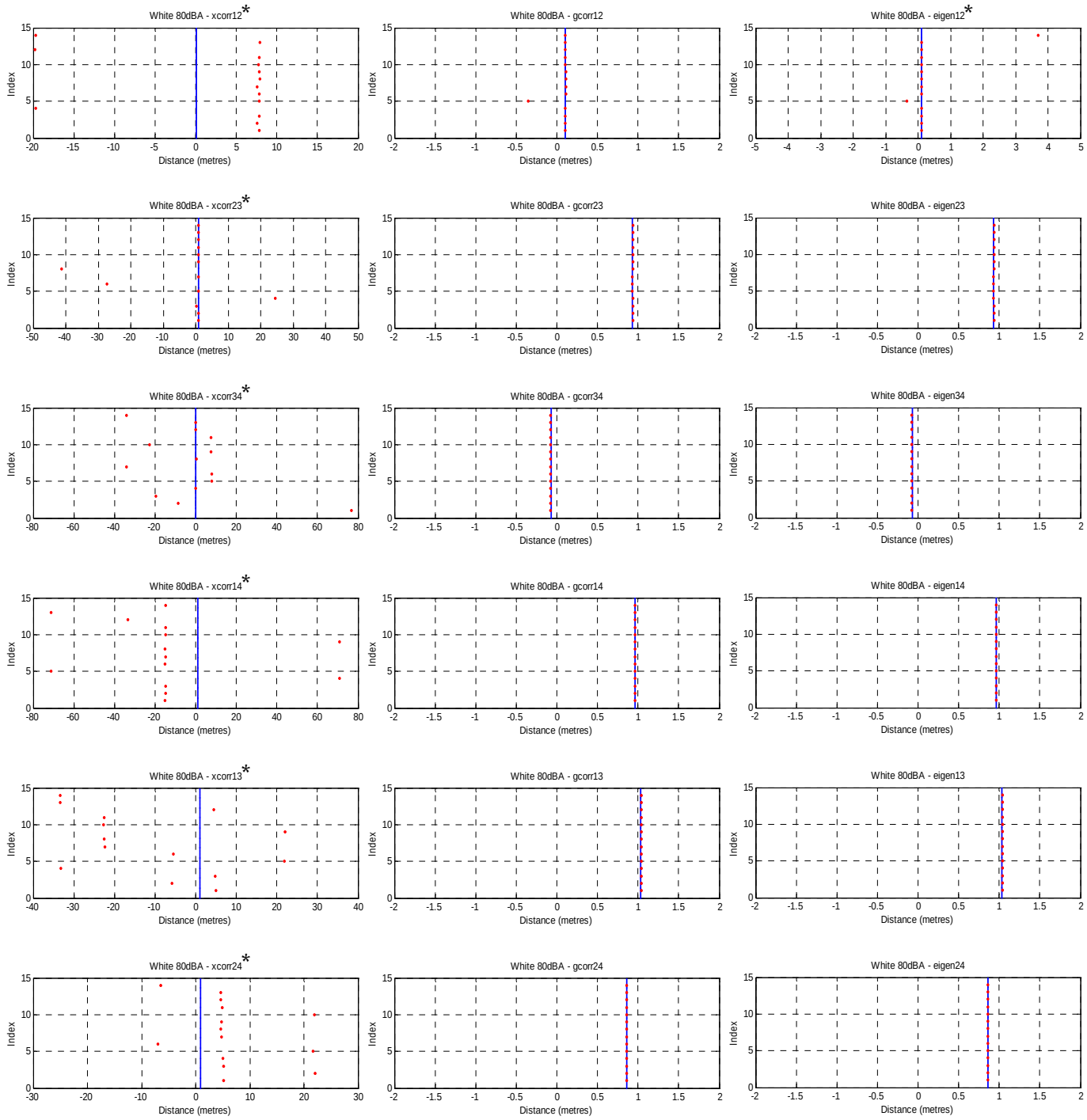

*Figure 94. Time-delay estimate plots using cross-correlation, GCC-PHAT and ED method for '80dBA' white noise. The red plots represent the estimates; the blue lines are the true time-delays; * indicates that the x-axis of that plot has a different scale.*

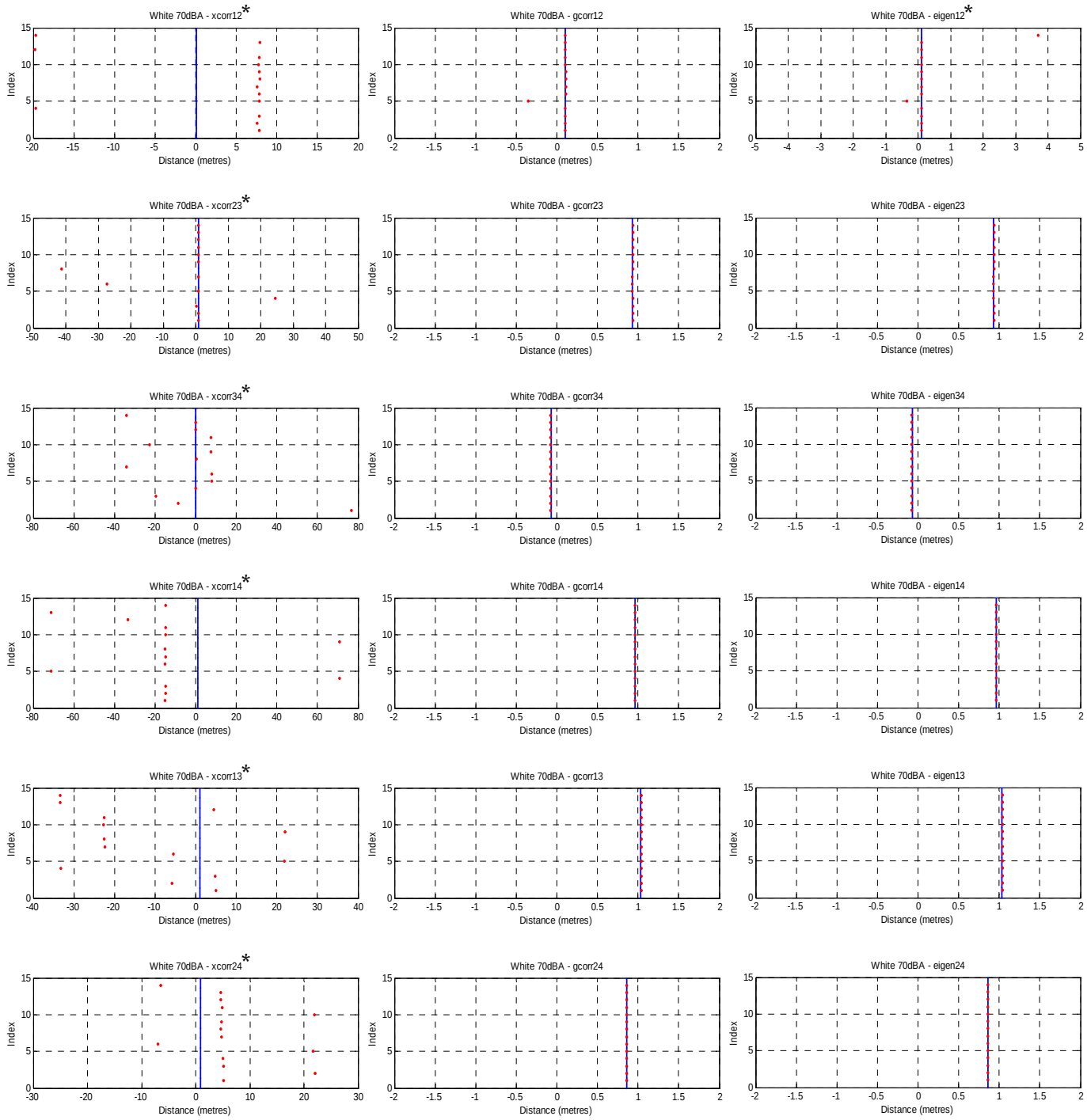

*Figure 95. Time-delay estimate plots using cross-correlation, GCC-PHAT and ED method for '70dBA' white noise. The red plots represent the estimates; the blue lines are the true time-delays; * indicates that the x-axis of that plot has a different scale.*

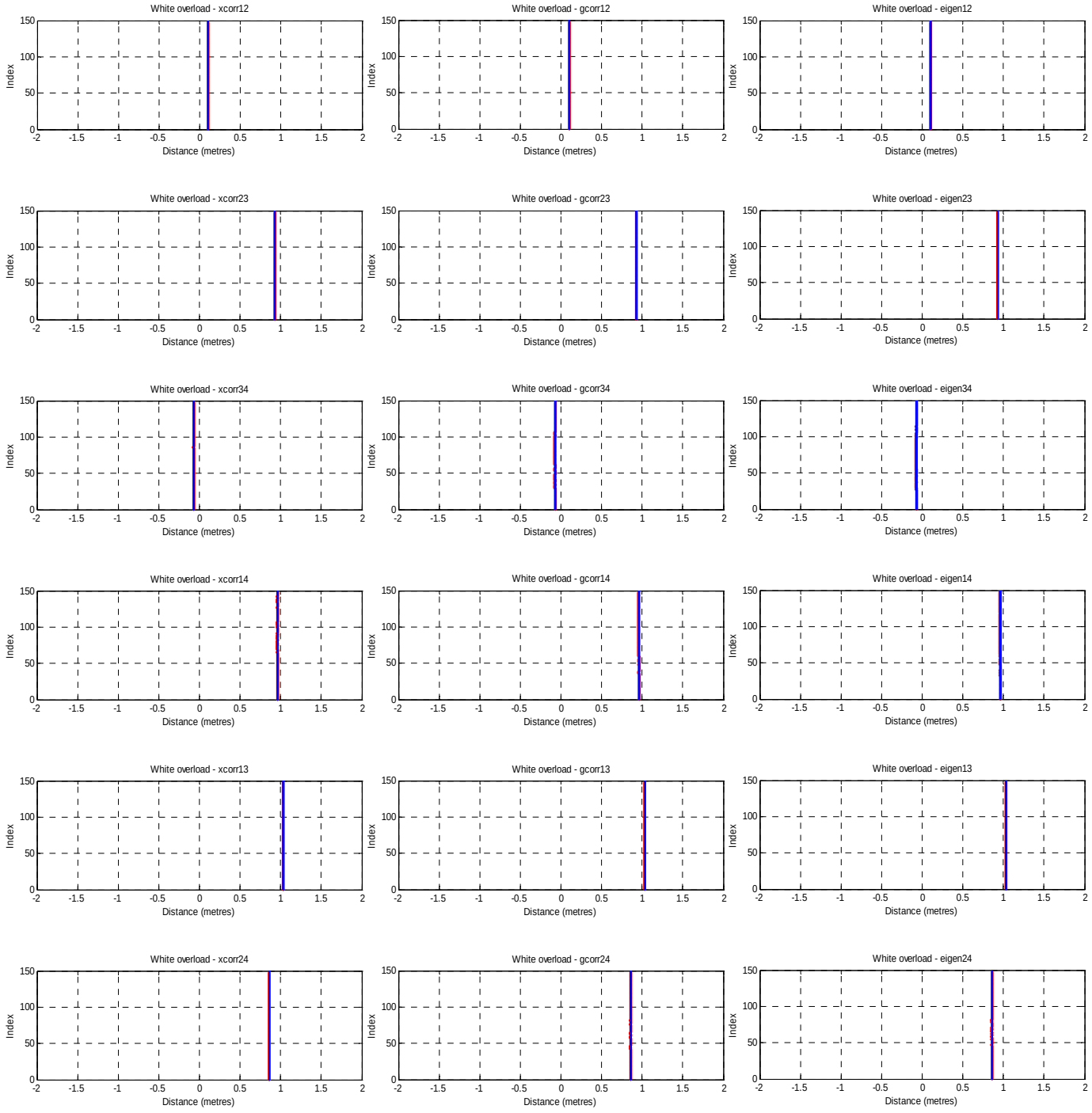

Figure 96. Time-delay estimate plots using cross-correlation, GCC-PHAT and ED method for 'overload' white noise. The red plots represent the estimates; the blue lines are the true time-delays.

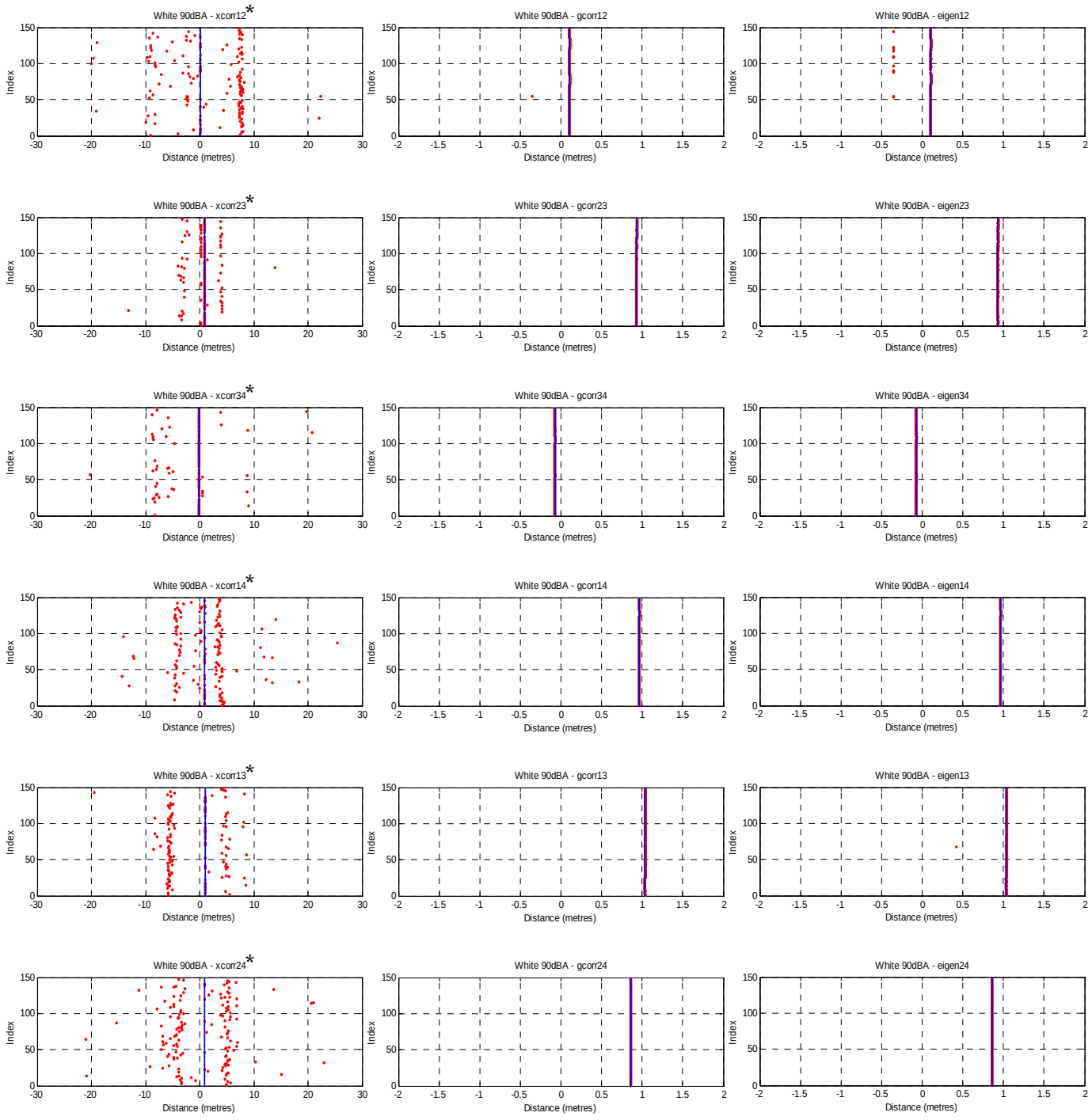

*Figure 97. Time-delay estimate plots using cross-correlation, GCC-PHAT and ED method for '90dBA' white noise. The red plots represent the estimates; the blue lines are the true time-delays; * indicates that the x-axis of that plot has a different scale.*

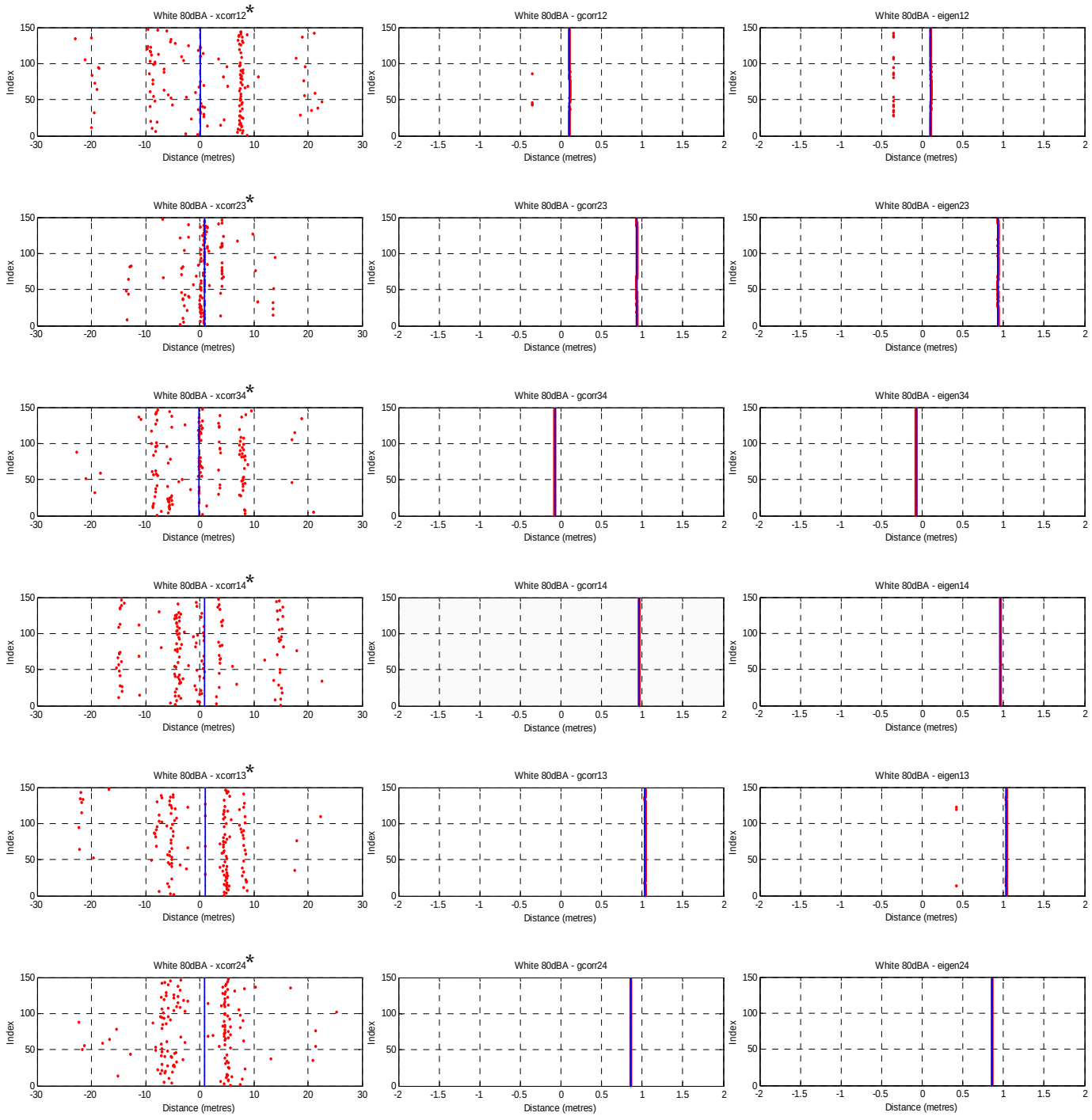

*Figure 98. Time-delay estimate plots using cross-correlation, GCC-PHAT and ED method for '80dBA' white noise. The red plots represent the estimates; the blue lines are the true time-delays; * indicates that the x-axis of that plot has a different scale.*

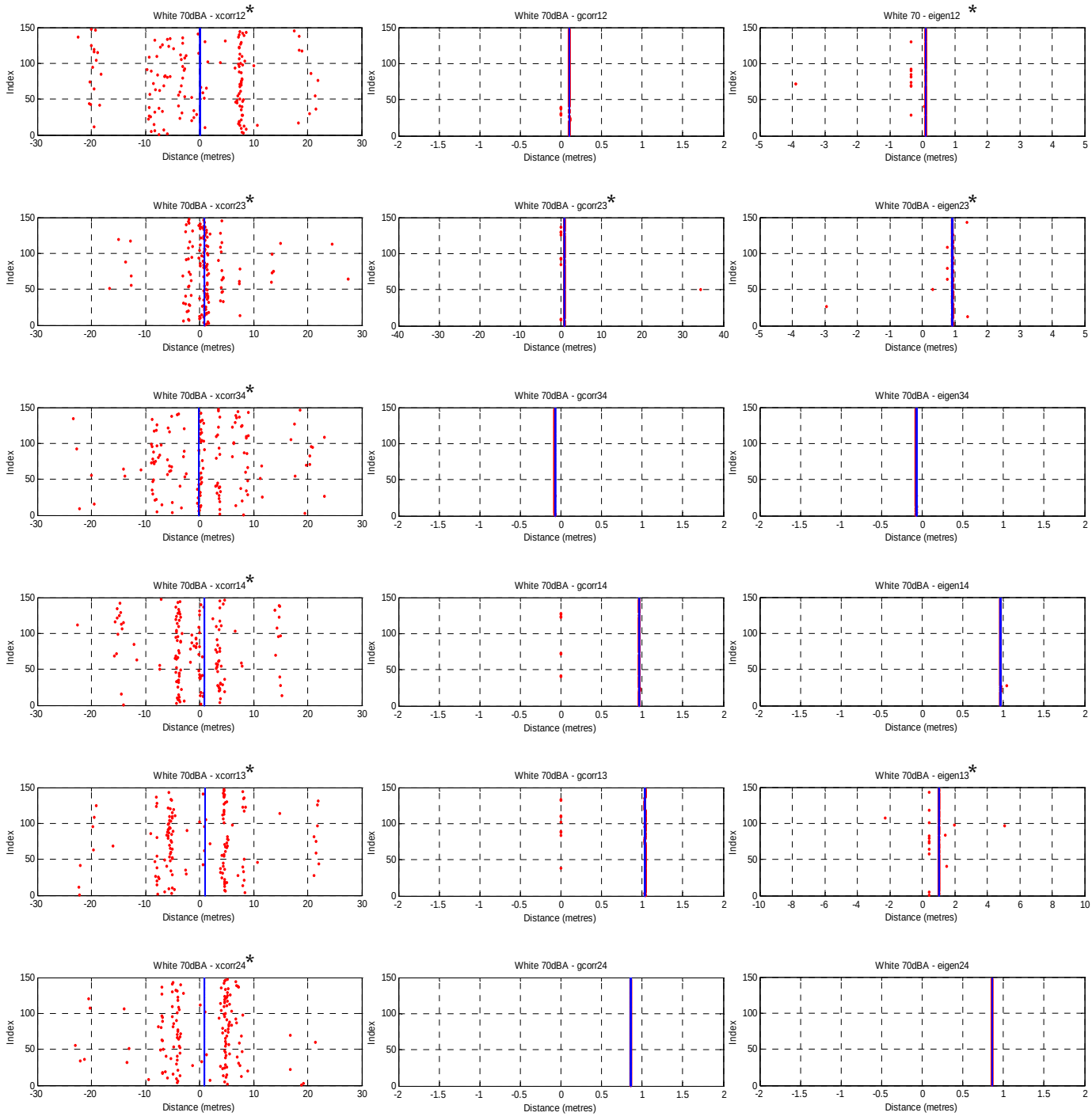

*Figure 99. Time-delay estimate plots using cross-correlation, GCC-PHAT and ED method for '70dBA' white noise. The red plots represent the estimates; the blue lines are the true time-delays; * indicates that the x-axis of that plot has a different scale.*

Appendix E: References

Ardoino, R., Capriati, F. & Zaccaron, A. (2006), 'Performances of a DTOA estimation algorithm based on Cross-Correlation Technique', *Proceedings of International Radar Symposium IRS 2006*, Kraków, Poland, 24-26 May 2006, pp. 189-192.

Bangs, W. & Schultheis, P. (1973), 'Space-time processing for optimal parameter estimation', *in Signal Processing (J. Griffiths, P. Stocklin & C. V. Schooneveld, eds.)*, Academic, New York, pp. 577-590.

Benesty, J. (2000), 'Adaptive eigenvalue decomposition algorithm for passive acoustic source localization', *Journal of the Acoustical Society of America*, Vol. 107, Issue 1, pp. 384-391.

Brandstein, M. S. & Silverman, H. F. (1997), 'A practical methodology for speech source localization with microphone arrays', Academic Press Limited.

Carter, G. (1977), 'Variance bounds for passively locating an acoustic source with a symmetric line array', *Journal of the Acoustical Society of America*, Vol. 62, pp. 922-926.

Chan, Y. T. & Ho, K. C. (1994), 'A Simple and Efficient Estimator for Hyperbolic Location', *IEEE Transactions on Signal Processing*, Vol. 42, No. 8.

Gatica-Perez, D., Lathoud, G., Odobez, J.-M and McCowan, I. (2006), 'Audio-visual probabilistic tracking of multiple speakers in meetings', *IEEE Transactions on Speech and Audio Processing*, February 2007. accepted for publication.

Gromov, K., Akos, D., Pullen, S., Eng, P. and Parkinson, B. (2000), 'GIDL: Generalized Interference Detection Localization System', *ION GPS 2000*, Department of Aeronautics and Astronautics, Stanford University.

Hahn, W. (1975), 'Optimum signal processing for passive sonar range and bearing estimation', *Journal of the Acoustical Society of America*, Vol. 58, pp. 201-207.

Hahn, W. & Tretter, S. (1973), 'Optimum processing for delay-vector estimation in passive signal arrays', *IEEE Transactions on Information Theory*, Vol. 19, pp. 608-614.

Haykin, S. (1991), *Adaptive Filter Theory*, Prentice Hall, Englewood Cliffs, NJ.

Johnson, D. & Dudgeon, D. (1993), *Array Signal Processing – Concepts and Techniques*, Prentice Hall, Englewood Cliffs, NJ.

Knapp, C. H. & Carter, G. C. (1976), 'The Generalized Correlation Method for Estimation of Time Delay', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-24, No. 4.

Macho, D., Padrell, J., Abad, A., Nadeu, C., Hernando, J., McDonough, J., Wölfel, M., Klee, U., Omologo, M., Brutti, A., Svaizer, P., Potamianos, G. & Chu, S. M. (2005), 'Automatic speech activity detection, source localisation, and speech recognition on the CHIL seminar corpus', *Proceedings of Int. Conf. Multimedia Expo.*, Amsterdam, The Netherlands.

Pang, D. (2006) 'Speaker Localisation and Tracking', *unpublished works as a Graduate Scholarship Student,* Command, Control, Communications and Intelligence Division, DSTO, Edinburgh.

Rice, F. (2004), 'Localising a Stationary Target by Using Time Difference of Arrival Information from Multiple UAVs', *Milestone I Report CR 01/04*, Sensor Signal Processing Group, School of Electrical and Electronic Engineering, The University of Adelaide.

Schau, H. C. & Robinson, A. Z. (1987), 'Passive Source Localization Employing Intersecting Spherical Surfaces from Time-of-Arrival Differences', *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-35, No. 8, pp. 1223-1225.

Silverman, H. F. & Kirtman, S. E. (1992), 'A two-stage algorithm for determining talker location from linear microphone-array data', *Computer Speech and Language*, Vol. 6, pp. 129-152.

Stein, S. (1981), 'Algorithms for Ambiguity Function Processing', *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-29, No. 3, pp. 588-599.

Wax, M. & Kailath, T. (1983), 'Optimum localization of multiple sources by passive arrays', *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-31, pp. 1210-1217.

Weber, D (2007) 'A Comparison of Livespace and Typical Meeting Room Support for Regular Software Development Review Meetings', *DSTO Technical Report, DSTO-TR-xxxx, Draft paper pending publication,* Command, Control, Communication and Intelligence Division, Defence Science and Technology Organisation, September 07.

Welsh, G. and Bishop, G. (1995), 'An Introduction to the Kalman Filter', *Technical Report TR 95-041*, Department of Computer Science, University of North Carolina.

Page classification: UNCLASSIFIED

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)
,

2. TITLE

Speaker Localisation using Time Difference of Arrival

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document (U)
Title (U)
Abstract (U)

4. AUTHOR(S)

Derek Zong Thai, Matthew Trinkle, Ahmad Hashemi-Sakhtsari, and Tim Pattison

5. CORPORATE AUTHOR

Command, Control, Communication and Intelligence Division, Defence Science and Technology Organisation, PO Box 1500, Edinburgh South Australia 5111 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-TR-2126	AR-014-178	Technical Report	April 2008

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
2007/1050658/2	LRR 07/248	CC3ID	97	22

13. URL on the World Wide Web

http://www.dsto.defence.gov.au/corporate/reports/DSTO-TR-2126.pdf

14. RELEASE AUTHORITY

Chief, C3I Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for Public Release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitation

17. CITATION IN OTHER DOCUMENTS Yes

18. DSTO RESEARCH LIBRARY THESAURUS http://web-vic.dsto.defence.gov.au/workareas/library/resources/dsto_thesaurus.htm

Algorithms
Auditory localisation
Speech processing
Speech recognition
Time difference of arrival

19. ABSTRACT

This report describes the research and development of speaker localisation to locate the position of a person speaking. Two closed-form localisation techniques were analysed, the first was developed by Schau and Robinson (1987) based on spherical intersection and the other developed by Chan and Ho (1994). Both techniques are based on time difference of arrival measurements. Accordingly three time delay estimators, namely cross-correlation, generalised cross-correlation, and an eigenvalue decomposition based algorithm were analysed. The implementation of the algorithms in Matlab and the results from the analyses are discussed.